ESD-TDR-63-474-1

MITRE-SS-1

FIRST CONGRESS ON THE INFORMATION SYSTEM SCIENCES

SESSION 1

CONCEPTS OF INFORMATION

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-63-474-1

FEBRUARY 1964

Prepared for

DIRECTORATE OF SYSTEM DESIGN
DEPUTY FOR TECHNOLOGY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

Project 704
Prepared by

THE MITRE CORPORATION
Bedford, Massachusetts
Contract AF33(600)-39852

FIRST CONGRESS ON THE INFORMATION SYSTEM SCIENCES

SESSION 1

CONCEPTS OF INFORMATION

TECHNICAL DOCUMENTARY REPORT NO. ESD-TDR-63-474-1

FEBRUARY 1964

Prepared for

DIRECTORATE OF SYSTEM DESIGN

DEPUTY FOR TECHNOLOGY

ELECTRONIC SYSTEMS DIVISION

AIR FORCE SYSTEMS COMMAND

UNITED STATES AIR FORCE

L. G. Hanscom Field, Bedford, Massachusetts

Project 704

Prepared by

THE MITRE CORPORATION

Bedford, Massachusetts

Contract AF33(600)-39852

CONCEPTS OF INFORMATION

Session Chairman: Rudolph F. Drenick

1. OUTLINES OF A FUTURE SYSTEMS THEORY
   Rudolph F. Drenick
2. A HEURISTIC DISCUSSION OF PROBABILISTIC DECODING
   Robert M. Fano
3. RECENT CONTROL SYSTEMS THEORY
   John G. Truxal

FIRST CONGRESS ON THE INFORMATION SYSTEM SCIENCES

Conducted at the Homestead, Hot Springs, Virginia

November 19—20, 1962

Chairman
Edward Bennett

Executive Manager for the MITRE Corporation
James Degan

Co-Chairman
Joseph Spiegel

Executive Manager for the Air Force Electronic Systems Division
Anthony Debons, Col., USAF

OUTLINES OF A FUTURE SYSTEMS THEORY

R. F. Drenick

ABSTRACT

This paper discusses the possibilities of evolving a systems theory which would take into account the important theories now existing. These include: Wiener's prediction and filtering theory; Shannon's information theory; control theory; automata theory; and others. A limited class of systems is specified to keep the discussion general and brief. "Noiseless," autonomous, noisy, and optimal systems are discussed, and the status of general theories for filtering, control, and communications systems, as well as additional systems problems, are described.

A HEURISTIC DISCUSSION OF PROBABILISTIC DECODING

R. M. Fano

ABSTRACT

This paper presents a heuristic discussion of the probabilistic decoding of digital messages after transmission through a randomly disturbed channel. The more general problem of transmitting digital information through randomly disturbed channels is outlined, and some of the key concepts and results pertaining to probabilistic decoding are reviewed.

A sequential decoding procedure recently developed by the author is described. The important characteristics of the procedure, i.e., complexity, the resulting probability of error per digit, and the probability of decoding failure are defined and described.

RECENT CONTROL SYSTEMS THEORY

J. G. Truxal

ABSTRACT

In this paper, the status of current control theory and research is discussed, with particular emphasis on the nature of the problems under consideration, the extent to which theory relates to engineering practice, and certain directions particularly promising for future developments. Those aspects of the theory which either have yielded interesting engineering results, or promise such a yield in the near future, are detailed.

The aspects of system design involved in "Stage 1," the transition from the customary, vague statement of broad system objectives to a configuration and a tractable model for the various elements of the system, are discussed initially. This is followed by a discussion of "Stage 2," which is concerned with the mathematical design of the "free" elements - the components in which at least certain of the parameters can be adjusted within specific bounds.

The two different approaches to optimization of system performance, restricted optimization and general optimization, are described. Recent work on the general problem of system performance with optimization based only on the specified process and signals is also described.

OUTLINES OF A FUTURE SYSTEMS THEORY

Rudolph F. Drenick*

SECTION I

INTRODUCTION

In the past few years, as has been variously observed,[7, 19] a new trend has developed in some of the theoretical work on systems. It is a trend towards the unification of systems concepts, a search for some of the basic features which all physical, or at least all man-made systems, have in common, and a study also of some of their most fundamental distinctions.

It is a very interesting field, and apparently also a fairly difficult one. If the term "inter-disciplinary" were not so widely abused, one would be tempted to use it here. For it is a matter of course that any future theory of systems must involve in some way the important theories which now exist, such as the prediction and filtering theory of Wiener, Shannon's information theory, control theory, automata theory, and others. What is more, the mathematical methods needed are by all indications considerably more advanced than any of those that been traditionally used in the field.

This article is intended to be a kind of status report. It is a very partial report, in both senses of the word "partial." For one, it is incomplete, and for another, it is biased. At this stage of developments, it seems difficult to avoid this. The trends have barely begun to emerge, or else have yet to take place. It is often a matter of purely personal opinion when one singles out

* Polytechnic Institute of Brooklyn

those trends one considers promising, or isolates areas which seem to be ready to burst into activity. Nevertheless, this is what is presented in this article. If all that can be said for it is that it is stimulating, that will have made it well worth writing.

SECTION II

GENERALITIES AND TERMINOLOGY

A system, in the sense in which the term is generally used in the information sciences and technology, is a device which can accept certain classes of physical quantities as inputs, and can generate certain other quantities as outputs. Systems can be, and have been classified in many ways: by their own physical nature and by those of their inputs and outputs; by whether they accept one, or more than one input, simultaneously, and whether they generate one or more than one output at the same time; by whether or not they change their charactertistics with time. One distinguishes linear and nonlinear systems, lumped and distributed parameter systems, active and passive systems, and so on. Terms like linear filters, transducers, finite and infinite machines, and many others have been used to single out special, and especially important, classes of systems.

In a general systems theory, one would evidently want to make as few distinctions as possible, at least at first, and treat as large a class of devices as possible. On the other hand, one must not go overboard on this either. If a class becomes all-embracing, chances are that nothing significant can be said about it. The striking of the proper balance is in fact a continuously exasperating dilemma in this field. One is always plagued with questions of whether or not the statements one can make about a large class of systems are still really "significant," or already so sweeping as to be virtually vacuous.

For the purpose of this article, we will restrict ourselves as follows: we shall deal only with

(i) stationary systems (also called time-invariant) whose characteristics do not change with time; and with

(ii) systems with either a single input and a single output, or two inputs and one output (and the latter only when we mention it).

Both restrictions are largely matters of convenience. They make our case easier to present, yet do not sacrifice very much in conceptual content.

It will be important in what follows to distinguish a few system types. For one, we will speak of continuous and of discrete systems. The former accept, and emit, a continuum of possible input values; for instance, all real numbers from -1 to +1, or from $-\infty$ to $+\infty$. The latter accept only a discrete set, often called an "alphabet" in this case; for instance, the integers 0 and 1, or all integers from 0 to ∞. Correspondingly, we will also speak of continuous signals and discrete signals.

A second distinction we will have to make is between singular and non-singular systems. A system is nonsingular, in Shannon's terminology, if there is a one-to-one relationship between every possible input signal to the system, and the output signal that results from that input. A nonsingular system, in other words, allows an input signal to be perfectly identified from the output signal, and vice versa. Singular devices, on the other hand, violate this property. Such violations can happen in many ways, but one which will matter, especially in what follows, is that due to noise in the system.

A nonsingular system evidently always has an (equally nonsingular) inverse; that is, a device which reconstructs the original system input from the output. A singular system on the other hand has no inverse.

A third distinction which will be made among systems is between causal and non-causal ones. A causal system is one whose output at any one time can depend only on the input at that time and at previous times. Causality is generally considered a basic property of all physical devices. It expresses the

fact that no such device can respond to a stimulus before it is received. A non-causal device, by contrast, can respond to, and hence can anticipate the future.

A notion related to causality, and the last to be mentioned here concerning systems, is the delay with which a system operates. It applies only to causal systems. If the output of such a system at any time depends only on the input, up to, say, one second prior to that time, the system is said to have the delay of one second. A system with a delay cannot have a causal inverse, even if it is nonsingular, since a precise reproduction (in time) of the original signal would require that the inverse also cancel the delay. This, however, it could only do by anticipation. The presence of delay in a system, therefore, raises the threat of non-causality.

Finally, we remark that we will treat most signals as random signals in our discussion below. This is in keeping, if nothing else, with a rather dominant tradition in work with systems.

SECTION III

THE "NOISELESS" COMMUNICATION THEORIES

It has been one of the curious side effects of the work on systems, rudimentary as it is, that some of the existing theories are now being viewed from a new perspective and that certain of their features appear to be gaining importance while others seem to be losing it. Among these theories are Wiener's prediction and filtering theory[16] and Shannon's information theory.[15] These are the ones which are referred to in the title of this section as the "communication theories."

Both theories fall rather naturally into two parts; one dealing with noiseless systems only, and the other with noisy ones. Shannon's theory, in fact, makes this distinction explicitly. Wiener does not, but his prediction theory can be viewed as dealing with noiseless systems. Both theories contain theorems of a kind which might be called "signal convertibility theorems." They are statements of conditions under which certain classes of signals are convertible into each other by means of appropriate systems. Neither theory, however, has given these particular theorems the prominence which, according to some present thinking, they deserve. It may, therefore, be useful to describe them here side by side, and to point out several similarities and distinctions.

While both theorems deal with random signals, Shannon's deals with discrete signals, while Wiener's deals with continuous ones. The probability measures must be stationary in both theories, which is a mild requirement. Otherwise, Shannon's theory places only some very mild restrictions on them, while in Wiener's theory they must be Gaussian, which is quite a drastic restriction (in principle anyway; in practice it is often quite acceptable).

Shannon's Theorem then states that any two discrete random signals can be converted into each other by a suitable system, and in fact, by a nonsingular one, provided only one condition is fulfilled. They must agree in a number which Shannon succeeded in attaching to every one of these random signals, namely their entropy. Wiener's theorem on the other hand deals with Gaussian random signals and states that any two of them can be converted into each other by an "essentially" nonsingular system of a very special kind, namely a linear filter. Moreover, this filter will be causal provided only another condition is fulfilled. This second condition is of a rather mathematical character, and it is often called the Paley-Wiener criterion (although it originated with Szegö). Both theorems, Shannon's as well as Wiener's, are constructive in that they not only assert the possibility of signal conversion, but also specify the devices which will accomplish it.

It may be noticed that Wiener's theorem assures us that the signal converters to which it leads are causal if the Paley-Wiener condition is fulfilled. In fact, it further assures us that if the condition is violated, either input signal or output signal will contain a curious ingredient, namely, an embedded signal which is perfectly predictable and hence not an altogether bona fide random signal. Wiener's theorem, we should add, ignores entropy altogether.

Shannon's theorem, on the other hand, ignores causality. In fact, the signal converters to which it leads require delays in general, often even infinite delays, so that their inverses are non-causal, as we have mentioned above. In other words, Shannon's requirement that the entropies be the same before and after conversion does not assure causality of the signal converter, in contrast to the Paley-Wiener criterion which does. The two conditions therefore have different purposes altogether.

These theorems carry many implications, one of which will be useful here. It is in the nature of a conceptual short-cut which is often used in mathematics. The theorems assert that the signals from certain classes are freely convertible into each other by suitable nonsingular systems. Such classes are then often called "equivalence classes," and some particular outstanding member of the class is chosen to represent the rest. Wiener's theorem can be looked on as lumping into one equivalence class all Gaussian signals without a perfectly predictable component. The equivalence relation in this case is the conversion by causal linear filters. The outstanding representative (despite some mathematically shady properties) is usually taken to be a random signal all of whose values are statistically independent. This signal is called white Gaussian noise. Shannon's theorem lumps into an equivalence class all signals with the same entropy and uses as equivalence relation the conversion by some appropriate system, causal or not. The outstanding representative is again a signal with all independent values (i.e., again a white noise through a discrete, not a Gaussian, one). Moreover, it is the signal whose values (alphabet letters) are equiprobable.

One of the basic results then of the two noiseless communication theories we have discussed here, and perhaps even the basic result of each, is the proof of the existence of such equivalence classes.

SECTION IV

STATUS OF A GENERAL THEORY OF NOISELESS SYSTEMS

There are curious similarities in Wiener's and Shannon's theories which virtually beg for an attempt at unification and extension. Such attempts have been in fact made recently, but they have been only partially successful at this writing.[18]

It stands to reason that, in a general theory of physical systems, the concept of causality should play a rather fundamental role. Causality, after all, is a fundamental property of physical systems. If this is so, it follows further that it is Wiener's theorem whose generalization should be attempted, and not Shannon's, for the latter ignores causality. Such a generalization might ideally say roughly this: "All random signals, Gaussian or not, fall into two classes; those that do, and those that do not, contain a perfectly predictable ingredient. Those that do not, are freely convertible into each other by causal nonsingular systems (linear or not), and thus can be freely generated from the reconverted to white noise."

If this theorem were true, it would constitute a sweeping generalization of Wiener's, and it would embrace Shannon's with room to spare. Unfortunately, it is certainly untrue. It has been shown by M. Rosenblatt[14] that there are random signals which cannot be generated from white noise by any nonsingular causal systems. Furthermore, there are also signals which cannot be reconverted to white noise by nonsingular causal devices.

In other words, there is no common market for random signals. The free convertibility a la Wiener is blocked by a class of nonconformist random signals. The full extent of the obstruction is not known, but one thing seems fairly certain.

It is the discrete signals, the ones figuring in Shannon's theorem and some of their relatives, which are the nonconformists.

Thus the ideal theorem we have proposed above must be modified. Three modifications suggest themselves. For one, the nonconformist signals can be excluded from it and treated in a separate theorem. For another, the requirement of system causality can be abandoned. Finally, the requirement of nonsingularity can be waived.

The question of course is whether or not these modifications do any good at all. The answer to this question is not known, though it is generally expected that it will be shown to be affirmative. In other words, it is believed that all three modifications can be carried out successfully, and it will be a matter of taste which is the most appealing. (The last one, however, is likely to win out, although it will often turn a noiseless system into a noisy one.) Partial results have been obtained by M. Rosenblatt, [14] Hansen, [9] and the writer, [6] those of Hansen being the most recent and most general.

There are hopes, at any rate, that sooner or later a very general theorem will become available which will make broad statements concerning the convertibility of signals by appropriate systems. Indications are, furthermore, that the statements will be constructive. They will assert not only the possibility of conversion, but will specify the nature of the signal converter. The latter will come in many forms, some already named (such as linear filters, finite-state machines, infinite-state machines, etc.) and others generally lumped under the nearly all-inclusive adjective "nonlinear."

The theorem will no doubt have a number of implications, among them one of possibly considerably consequence. Since it will make statements concerning infinite state machines, universal Turing machines among them, there seems

to be good reason for expecting some entanglement with the issues of computability and decidability. What forms this will take is unknown, but the mere prospect of it is most stimulating.

SECTION V

THE STRUCTURE OF NOISELESS SYSTEMS

In the preceding section, we have said of several theorems, some existing and some hoped for, that they are "constructive" because they specify the nature of certain signal converting systems. One may ask just how specific that specification really is, and how much of an indication these theorems give on how to construct such systems.

The answer unfortunately is that they do not give much of an indication. They specify the system usually in terms of a gigantic formula by which the output at any one time can be calculated if the input is known. The formula however does not in general give a clue on how a special device is to be constructed which will carry out the necessary calculations.

This is not considered a satisfactory situation, and some effort is under way towards improving it. The idea usually is to approximate, or even replace, a given complex formula by combinations of simpler ones. One feels (and often can also demonstrate) that the simpler formulae are also more readily reduced to practice.

One line of attack is based on a theorem by Cameron and Martin[4] and has been pursued with variations primarily by several workers, including Wiener.[17] This theorem, suitably interpreted, shows that a given system can under fairly general circumstances be built up from combinations of linear filters and memoryless nonlinear devices.

A second line of attack uses as its point of departure a Taylor series type of expansion for functionals, due originally to Volterra. It has since been variously extended, most recently by Balakrishnan[1] with the specific purpose of applying it to systems problems.

Both methods are what one might call brute-force methods. They can be used under an extremely wide range of circumstances, but they proceed fairly indiscriminately and can be quite inefficient. They are aimed mainly at continuous systems, but at least their most recent forms are applicable also to discrete systems. Finite-state systems in particular have been the subject of research along more specialized lines, and some methods of synthesis have been reported.[5, 11]

In neither case, as far as the present writer knows, do general synthesis procedures exist which even resemble those of linear networks. It is, of course, pointless to expect a great deal of resemblance here, but the absence of any is disturbing.

SECTION VI

AUTONOMOUS SYSTEMS

The systems we have considered so far accept a single input and convert it into a single output. These systems may be linear or not, but if they are linear and causal, one fact is well known concerning them: if the input is suddenly "turned off" (i.e., if an input which is constant and equal to zero is applied) the output from that moment on is characteristic of the system and allows highly revealing inferences to be drawn concerning its nature. In fact, knowing how a causal linear system behaves "autonomously", as the zero-input condition is often called, permits one to predict also its non-autonomous performance.

One can ask whether similar circumstances prevail with causal nonlinear systems. Can one predict the non-autonomous performance of such a system from its autonomous behavior?

The answer is a qualified yes. The qualification comes mainly from the fact that in nonlinear systems the condition of zero-input need not carry the same implications as in a linear one. With many such systems, constant non-zero inputs produce behaviors quite unlike that due to a constant-zero input, and hence in general all of them need to be considered. A given nonlinear system may, thus, have many associated autonomous systems, one for each possible constant input, and the behaviors of these various autonomous systems may differ considerably from each other.

It is easy to convince oneself, however, that if all autonomous behaviors of a system are known, then so is non-automonous performance.

This, unfortunately, is only the beginning of the problem, for it is next necessary to classify the possible autonomous behaviors and relate them to certain features, desirable or undersirable, of the non-autonomous system. Regarding this kind of problem, scarcely anything is known that is of much generality. The classification of autonomous behaviors could no doubt profit from two well-established mathematical disciplines, namely topological dynamics[13] and ergodic theory.[8] Considerable profit has in fact already been derived from the studies of the stable and unstable behavior of certain autonomous systems of the continuous type.

When it comes to the general problem of relating automonous and non-autonomous behaviors, no work exists to the writer's knowledge.

SECTION VII

NOISY SYSTEMS

In the preceding sections, we have discussed noiseless systems, and in fact mostly nonsingular noiseless systems, which have the property that to each input signal there is exactly one corresponding output signal, and vice versa. A noisy system is among those for which this is not so. Given an input signal, one can not be sure of the resulting output signal but can (in general) make only certain statistical statements concerning it; given the output signal, a similar uncertainty exists concerning the original input signal. (These two uncertainties usually do, but need not occur jointly. In many cases, it is the latter uncertainty which matters most.)

Noisy systems have been treated by Wiener as well as Shannon, though again only Shannon distinguished them explicitly from the noiseless ones. In Wiener's theory of prediction and filtering, it is the part dealing with filtering that involves noisy systems. One can ask whether or not a parallel can be established between Wiener's and Shannon's "noisy" theories, as was done for the noiseless theories in the preceding sections, and whether or not the parallel can be similarly exploited?

The answer to this question is not known. Indications, however, are that it is negative. Parallels, tempting as they are here, apparently do not exist. Shannon's and Wiener's noisy theories seem to be dealing with two quite dissimilar problems; Shannon's being probably much more difficult and unmanageable than Wiener's. Wiener's, on the other hand, may well be based on a more natural generalization of the concept of a noiseless system, or at any rate of the concept we have developed in this article. It is this generalization

which we shall now describe. (We shall have more to say about filtering theory a la Wiener in Sec. IX, and noisy communication theory a la Shannon in Sec. XI.)

In the preceding sections, we assigned fundamental importance in a noiseless communication theory to a theorem, as yet unproven, which stated in effect that, given any two random signals (preferably without perfectly predictable ingredients), a system could be found (preferably delay-free and nonsingular) which would accept one of the signals as input and generate the other as the output. We pointed out that this theorem, once proven, would subsume as special cases a theorem from Wiener's prediction theory, and another from Shannon's noiseless communication theory.

Now, in Wiener's filtering theory one can discover a theorem which has some points of resemblance with the one just mentioned, and which says the following: Any noisy linear system is equivalent to a system with two inputs, a main and a secondary one, and one output. The secondary input is a Gaussian noise input which can always be made statistically independent of the main one. If the main input and the output have no perfectly predictable ingredients, the system can be made causal and the noise white. Put in other words, any noisy linear system can for all purposes be replaced with another one in which all uncertainties and perturbations affecting the original are simulated by one independent (and often also white) Gaussian noise input. In many cases, the substitute system will even be causal.

There is reason for hope that this theorem can be substantially generalized. If it can, the new version will eliminate the words "linear" and "Gaussian." Like its noiseless counterpart, it will have trouble accommodating discrete signals, and the systems accordingly may not always come out causal. But in the main, the general theorem will no doubt sound very much like the special one from Wiener's theory. Moreover, since the proofs of the theorems from

Wiener's noiseless and noisy theories are very similar, one can hope the same for their generalizations. In other words, once the noiseless case is solved, the noisy one is likely to crack quickly.

Let us share in the optimism and assume in what follows that this theorem is proven.

SECTION VIII

OPTIMAL SYSTEMS

A large portion of future systems theory will no doubt deal with optimal systems. These are systems which are required to perform a certain task and to perform it better than (or at least as well as) any other system. The idea of such an optimum presumes two things, namely,

(i) that a class of systems is defined which are eligible for use and which may be entered into the competition for the optimum, and

(ii) that a criterion exists by which each system can be rated relative to all others, and a best one (or possibly be a best group) be selected from the rest.

Regarding item (i), the following can be said. In most problems concerning optimal systems, the competing systems are encumbered from the start by a severe handicap: a certain subsystem is given at the outset which must be accepted "as is," and all competing systems are required to incorporated it. The system designer, in other words, cannot negotiate over the given subsystem but must design the remainder of the system around it.

The given subsystem goes under various names (plant, object, channel, etc.) depending on which group of specialists deals with it, but it is always essentially the same thing. Let us call it the "plant" here. It may be a noiseless system, but in most problems it is a noisy one. The noise is often actually called noise, but sometimes also environment, uncertainty, etc. In any case, according to what we somewhat optimistically assumed in the preceding section, a noisy plant can then be viewed as a system with one output and two inputs, a main input and a noise input. What is more, it will often be possible to represent the latter by white noise, statistically independent

of the main input. This is illustrated in Fig. 1 in which y is the output, z the noise input, and x the main input.

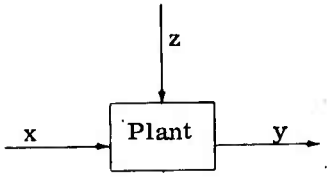


Fig. 1 A Noisy Subsystem

A system designer may thus assume that any plant with which he must contend in his design, is given in this form, along with certain specifications concerning the range of allowable inputs x, and concerning the uncertainty induced in y by the presence of z.

His problem will then usually consist of supplementing the plant with auxiliary apparatus which will ensure that the output of the complete system has certain desirable properties. The nature of the problem, however, and the method by which he will attack it, will depend very essentially on some other data of the problem, namely the point (or points) at which he may install the auxiliary apparatus and the signals which he may make available there.

Typically, three alternatives are open to him (which do not, however, exhaust all possible alternatives by any means). They are:

(a) insertion of a "filter" at the plant output with the idea of producing a more palatable system output u (see Fig. 2(a));

(b) insertion of a "controller" at the plant input with the idea of so conditioning x that the system output (in this case y) has the desired properties (in this arrangement, as shown in Fig. 2(b), the controller will often be supplied also with the system output y, making the system into a feedback system);

(c) insertion of a controller and a filter, or as they are more often called in this combination, of a coder and decoder (a feedback is sometimes provided for also in this arrangement, as shown in Fig. 2(c)).

Arrangement (a) leads to the so-called filtering problem. Arrangement (b) is the typical control system, especially when the feedback loop is present. Arrangement (c), minus the feedback path, is the traditional communication system. These three types are by now the garden varieties of systems, and they will be the only ones to be discussed in the section below, although they evidently are not the only ones that can be envisaged.

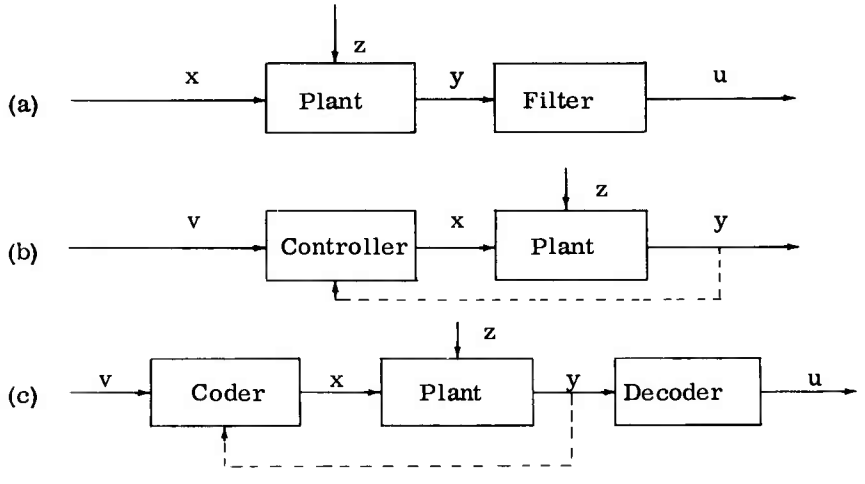


Fig. 2 "Garden Varieties" of Systems

First, however, it will be necessary to make some comments concerning item (ii) mentioned at the beginning of this section. It dealt with the need for a criterion of selection among the systems competing for the optimum. These systems, it has been said above, are required to impart to the output signal certain desirable characteristics. This can be visualized as meaning that a

prototype signal exists, in fact or in fiction, which actually has these characteristics, and that the performance of a system can be judged by how close its actual output comes to the prototype. In engineering language one would probably say that the system output is to "track" the prototype.

The question of relative system rating, on the basis of this argument, is equivalent to the definition of a measure of closeness, or of tracking performance, between output signal and prototype (random or deterministic). Many such measures have been proposed and used, the mean squared error criterion being the most popular one by far. This criterion penalizes a system at every instant by the square of the difference of the actual and desired system output at that instant. It rates the system by the time average of all instantaneous penalties.

Several generalizations of this are possible, and perhaps also desirable, and some have in fact been considered. For one, it is possible to admit penalties which, like the squared error, depend on the instantaneous values of input and output, but which involve these values in a way other than as the square of their difference. Such penalty functions go under various names, such as loss functions, fidelity evaluation functions, payoff functions, etc. Secondly, the penalty at any one time may be a function not only of the output values, actual and desired, at that same time, but also of the values at other times. The loss function, in other words, may be a loss functional. This possibility has been considered by Shannon. The loss function, or functional, may depend on quantities other than actual and desired output or, which is saying the same thing, the system may be assessed not only by how well it tracks the desired output but also by some additional criteria. Several such possibilities have been considered. Finally, system ratings may be considered which are not compounded from the individual ratings by mere averaging, or other linear operations, but by some nonlinear ones.

With this welter of possible criteria of selection, the question will then no doubt arise whether or not there is some way of singling out "reasonable" from "unreasonable" criteria. At this time, the viewpoint prevailing among systems workers is, that there is no such way. The performance criterion, it is argued, is a datum of the problem, provided to the system designer by some other agency, perhaps by an operations research group which has assessed the seriousness of an error or a failure to track the desired signals. This may very well be so. On the other hand, some evidence has begun to appear which indicates that at least under certain circumstances some criteria are more natural than others. The squared error criterion, for instance, is more natural than many others to linear problems when Gaussian random signals are involved. Similar associations are apparently possible between other criteria and nonlinear problems. This issue will be mentioned again below.

SECTION IX

STATUS OF A GENERAL FILTERING THEORY

The general filtering problem has been described briefly above, and illustrated in Fig. 2(a). Given is a noisy plant, to be followed by a filter. The latter is to be so designed that the system output y tracks a desired signal. Let this be the system input x (which is in fact often the case). We ignore the question of performance criterion for the moment.

A filter is in effect a decision-making device. It is presented, at every time t, with a certain amount of evidence, namely, the present and the past of the plant output y. On the basis of that, it is to make a judgment concerning the plant input x to which it has no direct access, and concerning which it can only make certain inferences. It is, more particularly, to generate an output, namely the system output u, whose value u_t at the time t constitutes the filter's best considered opinion, based on all available evidence, on what the value x_t of the system input is at that every same time.

If the plant is nonsingular, the filter's problem is clearly trivial since a nonsingular system by definition is one whose input can be perfectly reconstructed from its output. Hence, all the filter needs to do in this case is to carry out this reconstruction. It is true there may be trouble here. For instance, the filter may turn out non-causal and hence not strictly realizable. On the other hand, if the plant is noisy, and hence singular, a perfect reconstruction of the input is impossible, causally or otherwise. The problem of filtering is thus a genuine one.

There is one type of filtering problem in which the situation is well in hand, and this is the linear filtering problem, dealt with in Wiener's filtering theory. The plant in this problem is a noisy system in which a Gaussian

signal x and a Gaussian noise z are linearly superposed to generate an output y (again Gaussian). According to our comments in Sec. VII, we can mentally replace this plant with another, carrying the same input x into the same output y, but injecting a noise z which is independent of x. Let us assume we have done so.

Intuitively, one can then visualize the filter to have to proceed as follows. It is a decision-making device which is required to reconstruct x_t from the knowledge of the plant output y. The decision would be simple in principle, as we have just said, if it were not for the noise. On the other hand, the noise is statistically independent of x and hence cannot contribute any relevant information towards it. In making the decision, therefore, the filter should be justified in ignoring the noise because it is irrelevant; that is, it should treat the problem as if the noise were identically zero.

The remarkable thing is that this intuitive reasoning is correct, provided that the performance criterion is the mean squared-error. In other words, unless the loss function is of the squared-error (or some very similar) type, the filter cannot ignore in its decision what is, in a sense, irrelevant. This fact singles out the squared-error type of loss functions from many other possible ones as being naturally associated with Wiener's filtering problem.

It develops then further that the filter obtained by the mathematical version of this reasoning is itself linear, namely the inverse of the linear plant in which the noise has been set equal to zero. The filter is causal whenever y, the plant output, contains no perfectly predictable component.

In the nonlinear filtering problem, similar circumstances are likely to prevail though just how far the similarity will go is not known at this time. If the hopes expressed in Sec. VII come true, we will have at our disposal a theorem similar to Wiener's in which all noises affecting the plant can be

reduced to one, namely z, and more particularly to a noise which is statistically independent of the main input x. By our intuitive argument, statistical independence makes it irrelevant to the reconstruction of x from y, and should be immaterial to the filter design. We can hope that this intuitive argument can be made rigorous, as in the linear theory, by the choice of a suitable performance criterion. There is every reason to think that these things can be done and that a filtering theory can be evolved of greater generality than Wiener's but approaching his in elegance.

Until this happens, we can be content with a theory of lesser elegance but probably at least equal usefulness. Fortunately, we have such a theory. It ignores the issues of causality and perfect predictability, irrelevance and statistical independence, but it leads to filters of all kinds and for all kinds of purposes. It is based on the argument that the design of a decision-making device, such as a filter, should be based on a well-known statistical theory, established now for over a decade and developed for just such purposes. This is Wald's statistical decision theory, and more particularly, the easiest and least controversial part of it, namely the theory of Bayesian decisions. This point of view was taken by Middleton and van Meter in a paper[12] in which they showed that all standard filtering-type problems can be phrased as Bayesian decision problems. Once this was done, Wald's theory could be exploited for the production of filters by the score.

SECTION X

STATUS OF A GENERAL CONTROL THEORY FOR NOISY PLANTS

Control theory deals with the system problem illustrated in Fig. 2(b). It differs from the filtering problem in just one, apparently trivial, way; namely, in that the decision-making device (the controller) is required to precede the plant rather than to follow it. The difference, however, is far from trivial.

Let us assume here, as in the section above, that the system output y desired to be as close as possible (in some sense) to the system input, in this case v. Let us further assume that the plant is a causal system. This means that any input value x_t entered into it at the time t will have some effect on the present and future plant (and system) output y. The controller must then be an extremely "cautions" device, as decision-making devices go. Its decision to enter x_t is one whose consequences may have a bearing on all future system operation and hence on all future system performance. Such consequences cannot be shrugged off.

Note that the filters discussed in the preceding section, viewed as decision-making devices, needed to have no such worries. The decision at the time t to generate a certain system out u_t was made on the basis of all available evidence, namely the present and past plant output, but no thought needed to be given to any consequences of this decision on future system operation. The reason is simple: there were no such consequences.

In the control problem, however, there are. If controllers could be non-causal, if they could anticipate the future, they in fact could minimize their worries. The decision on what control signal to pick at any one time would admittedly still have to consider all consequences, but because of its unlimited

foresight, the controller could (impractical as this may be) chart all consequences precisely and finally select the most favorable plant input v_t. However, controllers typically are causal and cannot foretell the future, e.g., of the output of a noisy plant. Hence, this approach is not only impractical, it is impossible.

What is needed, therefore, is a technique for synthesizing optimal controllers, subject to the condition that they be causal. This technique exists. It is a recursion method, usually called "dynamic programming," which has been the subject of large numbers of articles and books, most of them including the authorship of Bellman.[2] As of this writing, however, not very much more can be said than that the technique exists and that it can, and often has been used, to synthesize controllers for certain specialized plants. Most of the controllers have been of the feedback type which is, in fact, the most important type. However, dynamic programming has not so far led to any new realizations which might be called "systems-theoretical" except in one area, and there they were, in fact, obtained by an interesting variant of dynamic programming due to Howard.[10] This is the area of finite-state systems.[3]

There is however, every reason to hope that further results along this line are imminent. Curiously, there is no indication so far that our much-heralded general theorem on the representation of noisy plants will be of great consequence here. This may lie in the nature of a control system, but more likely, is due to our limited understanding of the problem.

SECTION XI

THE STATUS OF A GENERAL THEORY OF NOISY COMMUNICATIONS SYSTEMS

The Communications system which plays the central role in the general theory of such systems is the one illustrated in Fig. 2(c). It is, of course, the system to which Shannon's theory is devoted, and more particularly, the portion dealing with noisy channels. This theory has led to the two fundamental coding theorems for noisy channels (or plants, as we have decided to call them in this article), one for discrete channels and the other for continuous ones. Together they constitute no doubt one of the greatest stimuli which the thinking in the natural sciences has received in this century.

If this appraisal is correct, one can ask what more there is that can conceivably be said in this area and that is not merely an elaboration of Shannon's ideas. At this stage, it seems impossible to produce more than some very circumstantial evidence which may indicate that Shannon's theorems do not answer all the basic questions that can be asked in this field.

Some of the pieces of evidence, such as they are, are as follows:

We have tried to make a case throughout this article that discrete systems and discrete signals apparently have some properties in which they differ fundamentally from their continuous counterparts. It is true Shannon's theory makes this distinction too, but in what seems a less fundamental way. In fact, when the showdown comes, namely when the main coding theorem is proved for the continuous case, the proof is carried out by reducing the continuous to the discrete channel. What evidence we now have seems to indicate that an important feature may have been lost in this process.

A related issue is concerned with the concepts of system causality and of entropy. We have seen that in the noiseless case, entropy played a significant role apparently only in the case of discrete signals and discrete systems. There was no need for it in the continuous problems. Yet in Shannon's theory of noisy channels, concepts related to the entropy are needed in the discrete as well as the continuous cases. This may well be as it should but the disparity in this respect between the noiseless and the noisy theories is a curious phenomenon.

A similar observation can be made concerning the concepts of causality and perfect predictability. We have seen that in the noiseless theory, non-causality in a system invariably (as far as we now know) injects a perfectly predictable component into the output. And once there, the component cannot be eliminated except by another non-causal system. Similar statements are possible also in filtering theory. Yet no such statement occurs in Shannon's theory.

Another curious point is this. The three diagrams of Fig. 2 show that a communications system is in a way a combination of a controller and a filter. However, there is preciously little resemblance in the theoretical treatment of communications system on one hand, and those of control systems and filters on the other. It is true that the basic statements of Shannon's coding theorems differ from the basic statements of the optimal control and filtering theories. Shannon's theorems assume a performance figure and state the conditions under which it can be reached or surpassed, but do not say now this can be done. The other two theories state the best performance figure that can be reached and say how it can be done. It seems reasonable to think that if one kind of statement can be made in one theory these similar statements ought to be possible in similar theories. However, no such statements have been made.

These are samples of the circumstantial evidence mentioned earlier which may suggest that not all of the fundamental issues have been raised in

communication theory. Perhaps, if they were raised, some additional insight could be gained also into the problems of coding and decoding which have preoccupied workers in this field for over a decade and which have so far proven most intractable to sweeping solutions.

SECTION XII

ADDITIONAL SYSTEMS PROBLEMS

This partial review of the outlines of a nonexistent systems theory would be altogether too partial if something were not said about several problems in this field which have been the subject of considerable discussion. Four of these problems will be mentioned briefly in this section. All of them, the present writer feels, share one common feature; namely, that despite much discussion, no clear trend has become discernible (at least to him) which might promise substantial and uncontroversial achievement in the near future.

The first of these is the field of adaptive or learning systems. As far as the writer knows, there exists no generally appealing, let alone a generally accepted, definition of when to call a system "adaptive" or "learning." Nor has it been decided whether the two terms are to be considered synonymous, and perhaps further synonymous also with "self-organizing." There is some indication that all three terms mean roughly the same thing to most persons, though control theorists perfer "adaptive" while computer people like "learning" better. Atempts at formalizing these concepts seem to have foundered on anthropomorphic objections and other preconceived notions.

In the present writer's opinion, these complications will shortly be overcome, and a widely acceptable theory of adaptation or learning will soon evolve. It is further his opinion, that the theory will not live up to the high expectations now held for it. Rather, it will develop to be essentially subsumed under the three theories discussed above.

The second area to be mentioned is the problem of uncertainty, as it is often called. It is argued that the theories we have described above are

houses built on sand, and the prettier the theory in many cases, the shakier its foundation. The reason for this is that they all assume a great deal of knowledge concerning the statistics of the random signals and the systems involved in all problems. Such knowledge typically is either not available or else not nearly as reliable as is generally assumed. Nor is the traditional gambit at all convincing which always replaces an uncertain parameter by a random variable because this replacement only introduces more parameters (namely those characterizing the probability distribution of the random variable) which are even more uncertain than the original one.

This complaint is very much to the point. On the other hand, it is probably also accurate to say that no good ideas exist on what to do about the situation in general. Statisticians in whose field of competence this problem falls seem as divided on how to deal with it as systems theorists. Perhaps the best thing to do is for the latter to wait and see what the former can agree on.

The third issue to be discussed here is what might be called the problem of the bad optimum. The complaint here is particularly against optimal systems theory, and it charges that the optima that come out of that theory (if they can be determined at all) are totally impractical. What is needed is a "workable sub-optimum."

This is a most reasonable request, but it is hard to fill. It has always been difficult to say when a theoretical solution was "workable," i.e., readily reducible to practice, and almost impossible to specify beforehand. The problem of how to find good and useful, rather than best solutions, has thus resisted precise formulation, and as far as the writer knows, none is in sight.

The last problem to be mentioned here is that of great complexity. The three canonical systems illustrated in Fig. 2 are patently gross oversimplifications. Systems of all kinds are much more complex in practice

as a rule, with potentially many subsystems in tandem or in parallel. The problem of synthesis, that is, of how to interconnect these many systems, and with what auxiliary equipment to supplement them, has received only little attention. Much the same is true of the problem of analysis, that is, of how to assess the performance of a given system of great complexity.

REFERENCES

1. A. V. Balakrishnan, A general theory of nonlinear estimation problem in control system, to be published.
2. e.g., R. Bellman, Dynamic Programming, Princeton University Press, 1957.
3. D. Blackwell, Discrete Dynamic Programming, Ann. Math. Stat. 33, (1962), 719-726.
4. R. H. Cameron and W. T. Martin, The orthogonal development of nonlinear functionals in series of Hermite polynomials, Ann. Math. 2, (1949),
5. I. M. Copi, C. L. Elgot, and J. B. Wright, Realization of events by logical nets, Jour. ACM 5, (1958), 181-192.
6. R. F. Drenick, On the Wald recomposition of non-Gaussian processes, (abstract), AMS Notices 8, (1961) 202.
7. M. M. Flood, New operations research potentials. Operations Research 10, (1962) 423-590.
8. R. R. Halmos, Lectures on Ergodic Theory, Publ's of Math. Soc., Japan, 1956.
9. D. L. Hansen, On the representation problem for stationary stochastic process with trivial tail field, (abstract), Bull AMS 68, (1962), 115-116.
10. R. Howard, Dynamic Programming and Markov Processes, John Wiley and Sons, 1960.
11. D. S. Huffman, Canonical forms for information-lossless finite-state logical machines, IRE Trans. CT-6 (1959), p. 41-59.
12. D. Middelton and D. van Meter, Detection and extraction of signals in noise from the standpoint of statistical decision theory, SIAM Jour. 3, (1955), 192-253, 4(1956), 86-119.
13. V. V. Nemitskii and V. V. Stepanov, Qualitative Theory of Differential Equations, Princeton University Press, 1960.
14. M. Rosenblatt, Stationary Processes as shifts of functions of independent random variables, Jour. Math. Mech. 8, (1959), 665-682.

15. C. E. Shannon and W. Weaver, A Mathematical Theory of Communications, Univ. Illinois Press, 1949.

16. N. Wiener, Extrapolation and interpolation of Stationary Time Series, John Wiley and Sons, 1946.

17. N. Wiener, Nonlinear Problems in Random Theory, John Wiley and Sons, 1958.

18. P. Masani and N. Wiener, Nonlinear prediction (article in U. Grenander, Ed., Probability and Statistics, John Wiley and Sons, 1959).

19. L. A. Zadeh, From circuit theory to systems theory, Proc. IRE 50 (1962), 856-866.

A HEURISTIC DISCUSSION OF PROBABILISTIC DECODING

by

Robert M. Fano*

SECTION I

INTRODUCTION

The purpose of this paper is to present a heuristic discussion of the probabilistic decoding of digital messages after transmission through a randomly disturbed channel. The adjective "probabilistic" is used to distinguish the decoding procedures discussed here from algebraic procedures [1] based on special structural properties of the set of code words employed for transmission.

In order to discuss probabilistic decoding in its proper frame of reference, we must first outline the more general problem of transmitting digital information through randomly disturbed channels, and review briefly some of the key concepts and results pertaining to it [2]. These key concepts and results were first presented by C. E. Shannon in 1948 [3], and later sharpened and extended by Shannon himself and others. The first probabilistic decoding procedure of practical interest was presented by J. M. Wozencraft in 1957 [4], and extended shortly thereafter by B. Reiffen [5]. Equipment implementing this procedure has been built at Lincoln Laboratory [6] and is presently being tested in conjunction with telephone lines.

---

*Ford Professor of Engineering, Massachusetts Institute of Technology. Department of Electrical Engineering and Research Laboratory of Electronics.
The work of this laboratory is supported in part by the U. S. Army Signal Corps, the Air Force Office of Scientific Research, and the Office of Naval Research.

SECTION II

THE ENCODING OPERATION

We shall assume, for the sake of simplicity, that the information to be transmitted consists of a sequence of equiprobable and statistically independent binary digits. We shall refer to these digits as information digits, and to their rate, R, measured in digits-per-second, as the information transmission rate.

The complex of available communication facilities will be referred to as the transmission channel. We shall assume that the channel can accept as input any time function whose spectrum lies within some specified frequency band, and whose r. m. s. value and/or peak value are within some specified limits.

The information digits are to be transformed into an appropriate channel input, and must be recovered from the channel output with as small a probability of error as possible. We shall refer to the device that transforms the information digits into the channel input as the encoder, and to the device that recovers them from the channel output as the decoder.

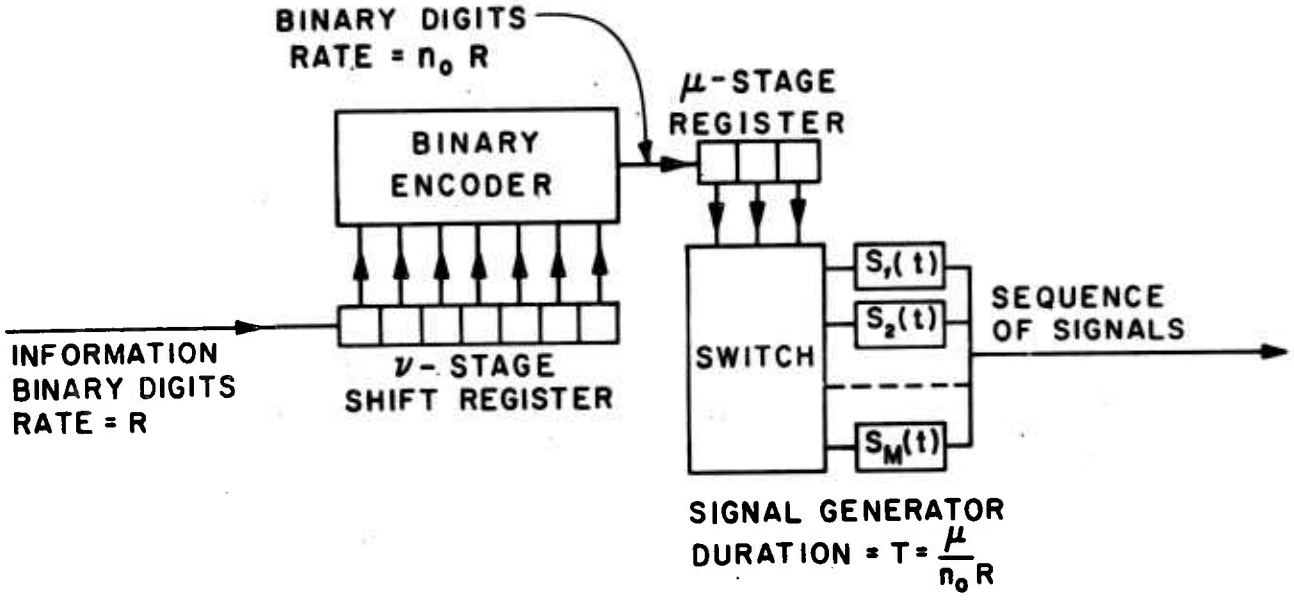
The encoder may be regarded, without any loss of generality, as a finite-state device whose state depends, at any given time, on the last ν information digits input to it. This does not imply that the state of the device is uniquely specified by the last ν digits. It may depend on time as well, provided such a time dependence is established beforehand and built into the decoder as well as into the encoder. The encoder output is uniquely specified by the current state, and therefore, is a function of the last ν information digits. We shall see that the integer ν, representing the number of digits on which the encoder output depends at any given time, is a critical parameter of the transmission process.

The encoder may operate in a variety of manners depending on how often ν digits are fed to it. The digits may be fed one at a time every 1/R seconds, or

two at a time every 2/R seconds, etc. The limiting case in which the information digits are fed to the encoder in blocks of ν every ν/R seconds is of special interest and corresponds to the mode of operation known as block encoding. In fact, if each successive block of ν digits is fed to the encoder in a time short compared to 1/R, the decoder output depends only on the digits of the last block and is totally independent of the digits of the preceding blocks. Thus, the encoder output during each time interval of length ν/R corresponding to the transmission of one particular block of digits is completely independent of the output during the time intervals corresponding to preceding blocks of digits. In other words, each block of ν digits is transmitted independently of all preceding blocks.

The situation is quite different when the information digits are fed to the encoder in blocks of size $\nu_o < \nu$. Then, the encoder output depends not only on the digits of the last block fed to the encoder, but also on $\nu - \nu_o$ digits of preceding blocks. Therefore, it is not independent of the output during the time interval corresponding to preceding blocks. As a matter of fact, a little thought will indicate that the dependence of the decoder output on its own past extends to infinity in spite of the fact that its dependence on the input digits is limited to the last ν. For this reason, the mode of operation corresponding to $\nu_o < \nu$ is known as sequential encoding. The distinction between block encoding and sequential encoding is basic to our discussion of probabilistic decoding.

The encoding operation, whether of the block or sequential type, is best performed in two steps, as illustrated in Fig. 1. The first step is performed by a binary encoder which generates no binary digits per input information digit, where the integer n_o is a design parameter to be selected in view of the rest of the encoding operation and of the channel characteristics. The binary encoder is a finite state device whose state depends on the last ν information digit fed to it, and possibly on time as discussed above. The dependence of the state on the



R = TRANSMISSION RATE IN BITS/SEC

ν, μ, n_0 = POSITIVE INTEGERS

S(t) = TIME FUNCTION OF DURATION T

M = NUMBER OF DISTINCT $S(t) \leq 2^{\mu}$

Fig. 1 The Encoding Operation

information digits is illustrated in Fig. 1 by showing the ν information digits as stored in a shift register with serial input and parallel output. It can be shown that the operation of the finite state encoder need not be more complex than a modular-2 convolution of the input digits with a periodic sequence of binary digits of period equal to $n_o\nu$. A suitable periodic sequence can be constructed by simply selecting the $n_o\nu$ digits equiprobably and independently at random. Thus, the complexity of the binary encoder grows linearly with ν, and its design depends on the transmission channel only through selection of the integers n_o and ν.

The second part of the encoding operation is a straightforward transformation of the sequence of binary digits generated by the binary encoder into a time function acceptable by the channel. Because of the finite state character of the encoding operation, the resulting time function must necessarily be a sequence of elementary time functions selected from a finite set. The elementary time functions are indicated in Fig.1 as $S_1(t)$, $S_2(t)$, . . . , $S_M(t)$, where M is the number of distinct elementary time functions and T is their common duration. The generation of these elementary time functions may be thought of as being controlled by a switch, whose position is in turn set by the digits stored in a μ-stage binary register. The digits generated by the binary encoder are fed to this register μ at a time, so that each successive group of μ digits is transformed into one of the elementary signals. The number of distinct elementary signals, M, can not exceed 2^μ, but it may be smaller. A value of M substantially smaller than 2^μ is used when some of the elementary signals are to be employed more often than others. For instance, with $M = 2$ and $\mu = 2$ we could make one of the two elementary signals occur three times as often as the other, by connecting three of the switch positions to one signal and the remaining one to the other.

While the character of the transformation of binary digits into signals envisioned in Fig. 1 is quite general, the range of the parameters involved is limited by practical considerations. The number of distinct elementary

signals, M, must be relatively small, and so must be the integer n_o. The values of M and n_o, as well as the forms of the elementary signals, must be selected with great care in view of the characteristics of the transmission channel. In fact, their selection results in the definition of the class of time functions that may be fed to the channel, and therefore, in effect, to a redefinition of the channel [7]. Thus, one faces here a compromise between equipment complexity and degradation of channel characteristics.

Fig. 2 illustrates two choices of parameters and of elementary signals, which would be generally appropriate when no bandwidth restriction is placed on the signal and thermal agitation noise is the only disturbance present in the channel. In case (a) each digit generated by the binary encoder is transformed into a binary pulse, while in case (b) each successive block of four digits is transformed into a sinusoidal pulse four times as long, and of frequency proportional to the binary number spelled by the group of four digits. The example illustrated in Fig. 3 pertains instead to the case in which the signal bandwidth is so limited that the shortest pulse duration permitted is equal to the time interval corresponding to the transmission of two information digits. In this case the elementary signals are pulses of the shortest permissible duration, with 16 different amplitudes.

These examples should make clear that the encoding process illustrated in Fig. 1 includes, as special cases, the traditional forms of modulation employed in digital communication. What distinguishes the forms of encoding envisioned here from the traditional forms of modulation is the order of magnitude of the integer ν. In the traditional forms of modulation the value of ν is very small, often equal to 1 and very seldom greater than 5. Here instead we envision values of ν of the order of 50 or more. The reason for using large values of ν will become evident later on.

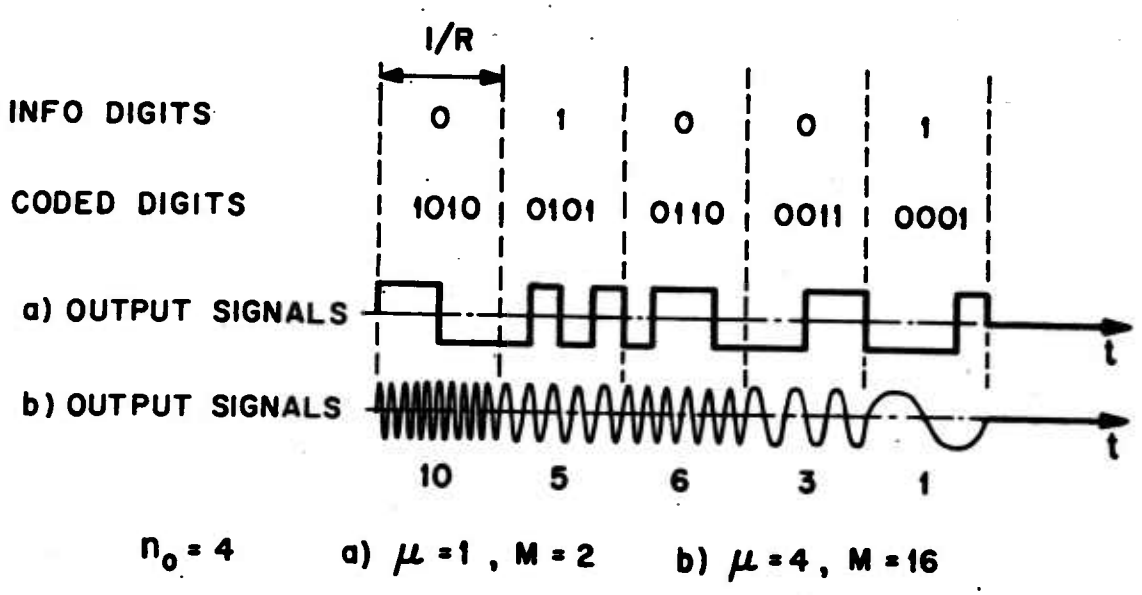


Fig. 2 Examples of Encoding for a Channel with Unlimited Band

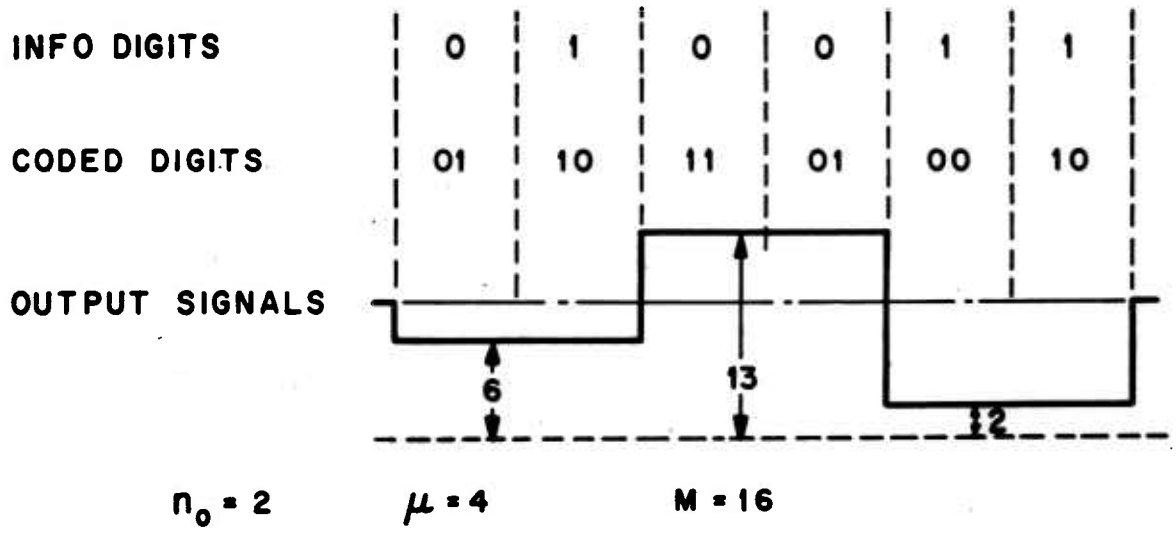


Fig. 3 Example of Encoding for a Band-Limited Channel

SECTION III

CHANNEL QUANTIZATION

Let us suppose that the encoding operation has been fixed to the extent of having selected the duration and identities of the elementary signals. We must consider next how to represent the effect of the channel disturbances on these signals. Since most of our present detailed theoretical knowledge is limited to channels without memory, we shall limit our discussion to such channels. A channel without memory can be defined for our purpose as one whose output during each time interval of length T, corresponding to the transmission of an elementary signal, is independent of the channel input and output during preceding time intervals. This implies that the operation of the channel can be described within any such time interval without reference to the past or the transmission. We shall assume as well that the channel is stationary in the sense that its properties do not change with time.

Let us suppose that the elementary signals are transmitted with probabilities $P(S_1)$, $P(S_2), \ldots, P(S_M)$, and indicate with $S'(t)$ the channel output during the time interval corresponding to the transmission of a particular signal. The observation of $S'(t)$ changes the probability distribution over the ensemble of elementary signals from the a priori distribution $P(S)$ to the a posteriori conditional distribution $P(S|S')$. The latter distribution can be computed, at least in principle, from the a priori distribution and the statistical characteristics of the channel disturbances. More precisely, we may regard the output $S'(t)$ as a point S' in a continuous space of suitable dimensionality. Then, if we indicate with $p(S'|S_k)$ the conditional probability density (assumed to exist) of the output S' for a particular input S_k, and with

$$p(S') = \sum_{k=1}^{M} P(S_k)\ p(S' \mid S_k)\ , \qquad (1)$$

the probability density of S' over all input signals, we have

$$P(S \mid S') = \frac{P(S)\ p(S' \mid S)}{p(S')}\ . \qquad (2)$$

Knowing the a posteriori probability distribution $P(S \mid S')$ is equivalent, for our purposes, to knowing the output signal S'. In turn, this probability distribution depends on S' only through the ratios of the M probability densities $p(S' \mid S)$. Furthermore, these probability densities can not be determined, in practice, with infinite precision. Thus, we must decide, either implicitly or explicitly, the tolerance within which the ratios of these probability densities are to be determined.

The effect of introducing such a tolerance is to lump together the output signals S' for which the ratios of the probability densities remain within the prescribed tolerance. Thus, we might as well divide the S' space into regions within which the ratios of the densities remain within the prescribed tolerance and record only the identity of the particular region to which the output signal S' belongs.

Such a quantization of the output space S' is governed by considerations similar to those governing the choice of the input elementary signals, namely equipment complexity and channel degradation. We shall not discuss this matter further, except for stressing again that such quantizations are unavoidable in practice, and that their net result is to substitute for the original transmission channel a new channel with discrete sets of possible inputs and outputs, and a correspondingly reduced transmission capability [7].

SECTION IV

CHANNEL CAPACITY

It is convenient at this point to change our terminology to that commonly employed in connection with discrete channels. We shall refer to the set of elementary input signals as the input alphabet and to the individual signals as input symbols. Similarly, we shall refer to the set of regions in which the channel output space has been divided as the output alphabet, and to the individual regions as output symbols. The input and output alphabets will be indicated with X and Y respectively, and particular symbols belonging to them will be indicated with x and y. Thus, the transmission channel is completely described by the alphabets X and Y, and by the set of conditional probability distributions $P(y|x)$.

We saw above that the net effect of the reception of a symbol y is to change the a priori probability distribution $P(x)$ into the a posteriori probability distribution

$$P(x|y) = \frac{P(x)\ P(y|x)}{P(y)} = \frac{P(x,y)}{P(y)}, \tag{3}$$

where $P(x,y)$ is the joint probability distribution of input and output symbols. Thus, the information provided by a particular output symbol y about a particular input symbol x is defined as

$$I(x;y) = \log \frac{P(x|y)}{P(x)} = \log \frac{P(y|x)}{P(y)} = \log \frac{P(x,y)}{P(x)\ P(y)}. \tag{4}$$

We shall see that this measure of information and its average value over the input and/or output alphabets play a central role in the problem under discussion.

It is interesting to note that $I(x;y)$ is a symmetrical function of x and y so that the information provided by a particular y about a particular x is the same as the information provided by x about y. In order to stress this symmetry property, $I(x;y) = I(y;x)$ is often referred to as the mutual information between x and y. By contrast,

$$I(x) = \log \frac{1}{P(x)} \tag{5}$$

is referred to as the self-information of x. This name follows from the fact that, for a particular symbol pair $x = x_k$, $y = y_i$, $I(x_k;y_i)$ becomes equal to $I(x_k)$ when $P(x_k|y_i) = 1$, that is when the output symbol y_i uniquely identifies x_k as the input symbol. Thus, $I(x_k)$ is the amount of information that must be provided about x_k in order to uniquely identify it, and as such is an upper bound to the value of $I(x_k;y)$.

In the particular case of an alphabet with L equiprobable symbols the self-information of each symbol is equal to log L. The information is measured in bits when base-2 logarithms are used in the above expressions. Thus, the self-information of the symbols of a binary equiprobable alphabet is equal to 1 bit.

Let us suppose that the input symbol is selected from the alphabet X with probability P(x). The average, or expected value, of the mutual information between input and output symbols is then,

$$I(X;Y) = \sum_{XY} P(x,y)\ I(x;y). \tag{6}$$

This quantity depends on the input probability distribution P(x) and on the characteristics of the channel represented by the conditional probability

distributions $P(y|x)$. Thus, its value, for a given channel, depends on the probability distribution $P(x)$ alone.

The channel capacity is defined as the maximum value of $I(X;Y)$ with respect to $P(x)$, that is

$$C = \underset{P(x)}{\text{Max}}\ I(X;Y) \tag{7}$$

It can be shown* that if a source which generates sequences of x symbols is connected to the channel input, the average amount of information per symbol provided by the channel output about the channel input can not exceed C, regardless of the statistical characteristics of the source.

*See Ref. 2, Sec. 5.2

SECTION V

ERROR PROBABILITY FOR BLOCK ENCODING

Let us consider now the special case of block encoding, and suppose that a block of ν information digits are transformed by the encoder into a sequence of N elementary signals, that is, into a sequence of N input symbols. Since the information digits are by assumption equiprobable and independent of one another, it takes an amount of information equal to log 2, (1 bit), to identify each of them. Thus, the information transmission rate per channel symbol is given by *

$$R = \frac{\nu}{N} \log 2. \tag{8}$$

The maximum amount of information per symbol that the channel output can provide about the channel input is equal to C, the channel capacity. It follows that we cannot expect to be able to transmit the information digits with any reasonable degree of accuracy at any rate $R > C$. Shannon's fundamental theorem asserts further that, for any $R < C$, the probability of erroneous decoding of a block of ν digits can be made as small as desired by employing a sufficiently large value of ν and a correspondingly large value of N. More precisely, it is possible ** to achieve a probability of error per block bounded by

$$P_e < 2^{-\nu \left(\frac{\alpha}{R} + 1 \right)}, \tag{9}$$

---

* The same symbol is used to indicate the information transmission rate, whether per channel symbol or per unit time.

** See Ref. 2, Ch. 9.

where α is independent of ν and varies with R as illustrated schematically in Fig. 4. Thus, for any $R < C$, the probability of error decreases exponentially with increasing ν.

It is clear from Eq. (9) that the probability of error is controlled primarily by the product of ν and α/R, the latter quantity being a function of R alone for a given channel. Thus, the same probability of error can be obtained with a small value of ν and relatively small value of R, or with a value of R close to C and a correspondingly larger value of ν. In the first situation, which corresponds to the traditional forms of modulation, the encoding and decoding equipment is relatively simple because of the small value of ν, but the channel is not utilized efficiently. In the second situation, on the contrary, the channel is efficiently utilized, but the relatively large value of ν implies that the terminal equipment must be substantially more complex. Thus, we are faced with a compromise between efficiency of channel utilization and complexity of terminal equipment.

It was pointed out in Section 1 that the operation to be performed by the binary encoder is relatively simple, namely the convolution of the input information digits with a periodic sequence of binary digits of period equal to $n_o \nu$. Thus, roughly speaking, the complexity of the encoding equipment grows linearly with ν. On the other hand, the decoding operation is substantially more complex both conceptually and in terms of the equipment required to perform it. The rest of this paper is devoted to it.

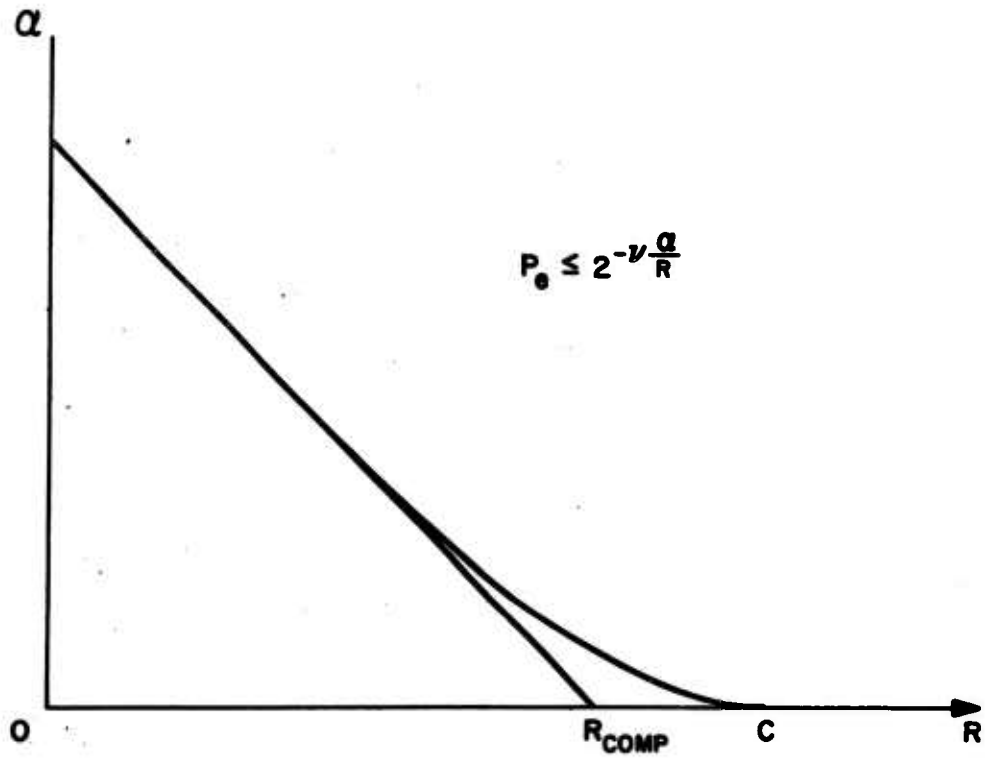


Fig. 4 Relation Between the Exponential Coefficient α and the Information Transmission Rate R in Eq. (9)

SECTION VI

PROBABILISTIC BLOCK DECODING

We saw in Section 1 that in the process of block encoding each particular sequence of ν information digits is transformed by the encoder into a particular sequence of N channel input symbols. We shall refer to any such sequence of input symbols as a codeword, and we shall indicate with u_k the codeword corresponding to the sequence of information digits which spells k in the binary number system. The sequence of N output symbols resulting from an input codeword will be indicated with v.

The probability that a particular codeword u will result in a particular output sequence v is given by

$$P(v|u) = \prod_{j=1}^{N} [P(y|x)]_j, \tag{10}$$

where the subscript j indicates that the value of the conditional probability is evaluated for the input and output symbols which occupy the j^{th} positions in u and v. On the other hand, since all sequences of information digits are transmitted with the same probability, the a posteriori probability of any particular codeword u after the reception of a particular output sequence v is given by

$$P(u|v) = \frac{P(v|u)\,P(u)}{\sum_{U} P(v|u)\,P(u)} = 2^{-\nu}\,\frac{P(v|u)}{P(v)} . \tag{11}$$

Thus, the codeword which is a posteriori most probable for a particular output v is the one that maximizes the conditional probability $P(v|u)$ given by Eq. (10). We can conclude that, in order to minimize the probability of error,

the decoder should select the codeword with the largest probability, $P(v|u)$, of generating the sequences v output from the channel.

While the specification of the optimum decoding procedure is straightforward, its implementation presents very serious difficulties for any sizable value of ν. In fact, there is no general procedure for determining the codeword corresponding to the largest value of $P(v|u)$ without having to evaluate this probability for most of the 2^{ν} possible codewords. Clearly, the necessary amount of computation grows exponentially with ν and becomes prohibitively large very quickly. However, if we do not insist on minimizing the probability of error, we may take advantage of the fact that, if the probability of error is to be very small, the a posteriori most probable codeword must be almost always substantially more probable than all other codewords. Thus, it may be sufficient to search for a codeword with a value of $P(v|u)$ larger than some appropriate threshold and take a chance on the possibility that there be other codewords with even larger values, or that the value for the correct codeword be smaller than the threshold.

Let us consider then what might be an appropriate threshold. Let us suppose that, for a given received sequence v, there exists a codeword u_k for which

$$P(u_k|v) \geq \sum_{i \neq k} P(u_i|v), \tag{12}$$

where the summation extends over all the other $2^{\nu}-1$ codewords. Then u_k must be the a posteriori most probable codeword. The condition expressed by Eq. (12) can be rewritten, with the help of Eq. (11), as

$$P(v|u_k) \geq \sum_{i \neq k} P(v|u_i). \tag{13}$$

The value of $P(v|u_k)$ can be readily computed with the help of Eq. (10). However, we are still faced with the problem of evaluating the same conditional probability for all the other codewords. This difficulty can be circumvented by using an approximation related to the random-coding procedure employed in deriving Eq. (9).

In the process of random coding each codeword is constructed by selecting its symbols independently at random according to some appropriate probability distribution $P_o(x)$. The right-hand side of Eq. (9) is actually the average value of the probability of error over the ensemble of codeword sets so constructed. This implies, incidentally, that satisfactory codewords can be obtained in practice by following such a random construction procedure.

Let us assume that the codewords under consideration have been constructed by selecting the symbols independently at random according to some appropriate probability distribution $P_o(x)$. It would seem reasonable then to substitute for the right-hand side of Eq. (13) its average value over the ensemble of codeword sets constructed in the same random manner. In such an ensemble of codeword sets, thc probability $P_o(u)$ that any particular input sequence u be chosen as a codeword is

$$P_o(u) = \prod_{j=1}^{N} \left[P_o(x)\right]_j , \tag{14}$$

where the subscript j indicates that $P_o(x)$ is evaluated for the j^{th} symbol of the sequence u. Thus, the average value of the right-hand side of Eq. (12) is, with the help of Eq. (10),

$$(2^{\nu}-1) \sum_{U} P_o(u)\, P(v|u) = (2^{\nu}-1) \prod_{j=1}^{N} \left[P_o(y)\right]_j , \tag{15}$$

where U is the set of all possible input sequences, and

$$P_o(y) = \sum_X P_o(x)\ P(y|x) \tag{16}$$

is the probability distribution of the output symbols when the input symbols are transmitted independently with probability $P_o(x)$. Then, substituting the right-hand side of Eq. (15) for the right-hand side of Eq. (13), and expressing $P(v|u_k)$ as in Eq. (10), yields

$$\prod_{j=1}^{N} \left[\frac{P(y|x)}{P_o(y)}\right]_j \geq 2^{\nu}-1. \tag{17}$$

Finally, approximating $2^{\nu}-1$ with 2^{ν} and taking the logarithm of both sides yields

$$\sum_{j=1}^{N} \left[\log \frac{P(y|x)}{P_o(y)}\right]_j \geq N\ R\ , \tag{18}$$

where R is the transmission rate per channel symbol defined by Eq. (8).

The threshold condition expressed by Eq. (18) can be given a very interesting interpretation. The j^{th} term of the summation is the mutual information between the j^{th} output symbol and the j^{th} input symbol, with the input symbols assumed to occur with probability $P_o(x)$. If the input symbols were statistically independent of one another, the sum of these mutual informations would be equal to the mutual information between the output sequence and the input sequence. Thus, Eq. (18) states that the channel output can be safely decoded into a particular codeword if the mutual information that it provides about the codeword, evaluated as if the N input symbols were selected independently with probability $P_o(x)$, exceeds the amount of information transmitted per codeword.

It turns out that the threshold value on the right-hand side of Eq. (18) is not only a reasonable one, as indicated by our heuristic arguments, but the one which minimizes the average probability of error for threshold decoding over the ensemble of randomly constructed codeword sets. This was shown by C. E. Shannon in an unpublished memorandum. The bound on the probability of error obtained by Shannon is of the form of Eq. (9); however, the value of α is somewhat smaller than that obtained for optimum decoding. Shannon assumes in his derivation that an error occurs whenever Eq. (17) is satisfied for any codeword other than the correct one and whenever it is not satisfied for the correct codeword.

The fact that the probability of error for threshold decoding, although larger than for optimum decoding, is still bounded as in Eq. (9), encourages us to look for a search procedure that will quickly reject any codeword for which Eq. (17) is not satisfied and thus converge relatively quickly on the codeword actually transmitted. We observe, on the other hand, that, even if we could reject an incorrect codeword after evaluating Eq. (17) over some small but finite fraction of the N symbols, we would still be faced with an amount of computation that would grow exponentially with ν. In order to avoid this exponential growth we must arrange matters in such a way as to be able to eliminate large subsets of codewords by evaluating the left-hand side of Eq. (17) over some fraction of a single codeword. This implies that the codewords must possess the kind of tree structure that results from sequential encoding, as discussed in the next section.

It is just the realization of this fact that led J. M. Wozencraft to the development of his sequential decoding procedure in 1957. Other decoding procedures, both algebraic [1] and probabilistic [8], have been developed since, which are of practical value in certain special cases. However, sequential decoding remains the only known procedure which is applicable to all channels without memory. As a matter of fact, there is reason [9] to believe that some modified form of sequential decoding may yield satisfactory results in conjunction with a much broader class of channels.

SECTION VII

SEQUENTIAL DECODING

The rest of this paper is devoted to a heuristic discussion of a sequential decoding procedure recently developed by the author. This procedure is similar in many respects to that of Wozencraft [4,5,6], but it is conceptually simpler and therefore it can be more readily explained and evaluated analytically. An experimental comparison of the two procedures is in progress at Lincoln Laboratory. A detailed analysis of the newer procedure will be presented in a forthcoming paper.

Let us reconsider in greater detail the structure of the encoder output in the case of sequential encoding, that is when the information digits are fed to the encoder in blocks of size ν_o (in practice ν_o is seldom larger than 3 or 4). The encoder output, during the time interval corresponding to a particular block, is selected by the digits of the block from a set of 2^{ν_o} distinct sequences of channel input symbols. The particular set of sequences from which the output is selected is specified, in turn, by the $\nu - \nu_o$ information digits preceding the block in question. Thus, the set of possible outputs from the encoder can be represented by means of a tree with 2^{ν_o} branches stemming from each node. Each successive block of ν_o information digits causes the encoder to move from one node to the next one along the branch specified by the digits of the block.

The two trees shown in Fig. 5 correspond to the two examples illustrated in Fig. 2-b and Fig. 3. The first example yields a binary tree ($\nu_o = 1$), while the second example yields a quaternary tree ($\nu_o = 4$).

In summary, the encoding operation can be represented in terms of a tree in which the information digits select at each node the branch to be followed. The path in the tree resulting from the successive selections constitutes the encoder

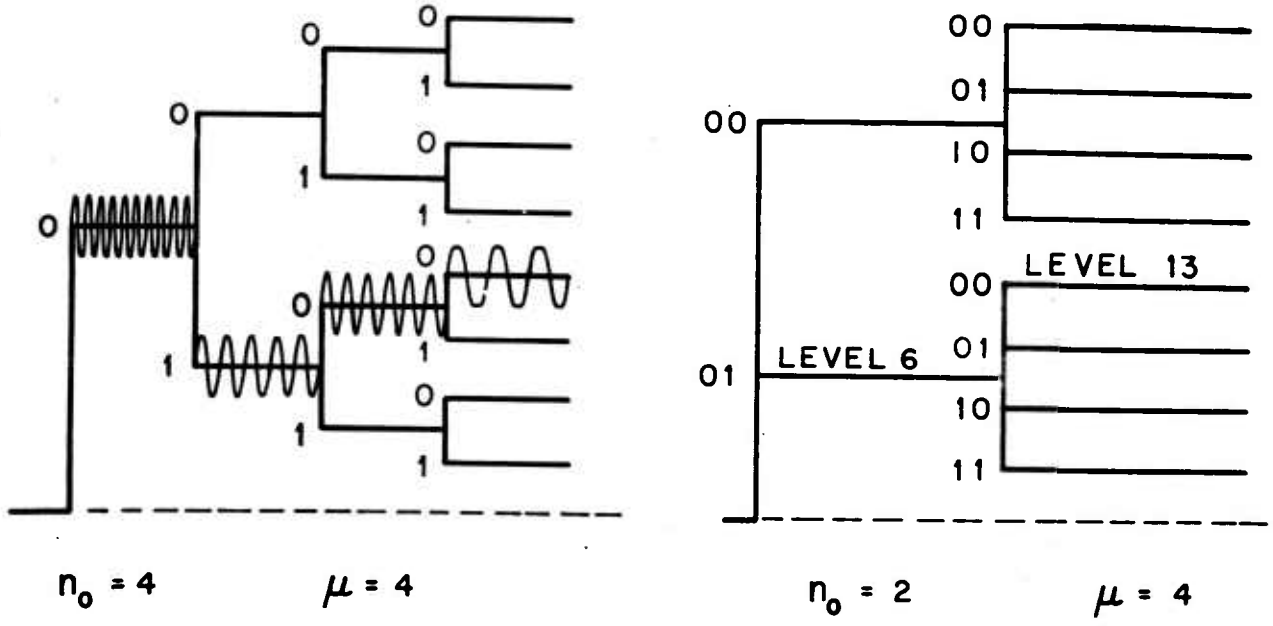


Fig. 5 Encoding Trees Corresponding to the Examples of Figs. 2(b) and 3

output. This is equivalent to saying that each block of ν_0 digits fed to the encoder is represented for transmission by a sequence of symbols selected from a set of 2^{ν_0} distinct sequences, but the particular set from which the sequence is selected depends on the preceding $\nu - \nu_0$ information digits. Thus, the channel output during the time interval corresponding to a block of ν_0 information digits provides information not only about these digits but also about the preceding $\nu - \nu_0$ digits.

The decoding operation may be regarded as the process of determining from the channel output the path in the tree followed by the encoder. Suppose, to start with, that the decoder selects at each node the branch which is a posteriori most probable on the basis of the channel output during the time interval corresponding to the transmission of the branch. If the channel disturbance is such that the branch actually transmitted does not turn out to be the most probable one, the decoder will make an error thereby reaching a node which does not lie on the path followed by the encoder. Thus, none of the branches stemming from it will appear as a likely channel input. If by accident one branch does appear as a likely input, the same situation will arise with respect to the branches stemming from the node in which it terminates, and so forth and so on. This rough notion can be made more precise as follows.

Let us suppose that the branches of the tree are constructed, as in the case of block encoding, by selecting symbols independently at random according to some appropriate probability distribution $P_0(x)$. This is accomplished in practice by selecting equiprobably at random the $n_0\nu$ binary digits specifying the periodic sequence with which the sequence of information digits is convolved, and by properly arranging the connections of switch positions to elementary signals in Fig. 1. Then, as in the case of threshold block decoding, the decoder, as it moves along a path in the tree, evaluates the quantity

$$I_N = \sum_{j=1}^{N} \left[\log \frac{P(y|x)}{P_o(y)} \right]_j , \qquad (19)$$

where y in the j^{th} term of the summation is the j^{th} symbol output from the channel, and x in the same term is the j^{th} symbol along the path followed by the decoder.

As long as the path followed by the decoder coincides with that followed by the encoder I_N can be expected to remain greater than N R. (R, the information transmission rate per channel symbol, is still equal to the number of channel symbols divided by the number of corresponding information digits, but it is no longer given by Eq. (8).) However, once the decoder has made a mistake and has thereby arrived to a node which does not lie on the path followed by the encoder, the terms of I_N corresponding to branches beyond that node are very likely to be smaller than R. Thus I_N must eventually become smaller than NR, thereby indicating that an error must have occurred at some preceding node. It is clear that in such a situation the decoder should try to find the place where the mistake has occurred so as to get back on the correct path. It would be desirable therefore to evaluate for each node the relative probability that a mistake has occurred there.

SECTION VIII

PROBABILITY OF ERROR ALONG A PATH

Let us indicate with N the order number of the symbol preceding some particular node, and with N_o the order number of the last output symbol. Since all paths in the tree are a priori equiprobable, their a posteriori probabilities are proportional to the conditional probabilities $P(v|u)$ where u is the sequence of symbols corresponding to a particular path, and v is the resulting sequence of output symbols. This conditional probability can be written in the form

$$P(v|u) = \prod_{j=1}^{N} [P(y|x)]_j \prod_{j=N+1}^{N_o} [P(y|x)]_j . \qquad (20)$$

The first factor on the right-hand side of Eq. (20) has the same value for all the paths which coincide over the first N symbols. The number of such paths, which differ in some of the remaining $N_o - N$ symbols, is,

$$m = 2^{(N_o - N)R/\log 2} . \qquad (21)$$

As in the case of block decoding, it is impractical to compute the second factor on the right-hand side of Eq. (20) for each of these paths. We shall again circumvent this difficulty by averaging over the ensemble of randomly constructed trees. By analogy with the case of threshold block decoding we obtain

$$P_o(v|u) = \prod_{j=1}^{N} [P(y|x)]_j \prod_{j=N+1}^{N_o} [P_o(y)]_j , \qquad (22)$$

where $P_o(y)$ is given by Eq. (16).

Let P_N be the probability that the path followed by the encoder is one of the m-1 paths which coincide with the one followed by the decoder over the first N symbols, but differ from it in some of the remaining symbols. We have, by approximating m-1 with m,

$$P_n = K_1 \, 2^{(n_o - N)R/\log 2} \prod_{j=1}^{N} [P(y|x)]_j \prod_{j=N+1}^{N_o} [P_o(y)]_j \qquad (23)$$

$$= K_2 \, 2^{-N R/\log 2} \prod_{j=1}^{N} \left[\frac{P(y|x)}{P_o(y)}\right]_j ,$$

where K_1 and K_2 are proportionality constants. Finally, taking the logarithm of both sides of Eq. (23) yields

$$\log P_N = \log K_2 + \sum_{j=1}^{N} \left[\log \frac{P(y|x)}{P_o(y)} - R\right]_j . \qquad (24)$$

The significance of Eq. (24) is best discussed after rewriting it in terms of the order number of the nodes along the path followed by the decoder. Let us indicate with N_b the number of channel symbols per branch (assumed for the sake of simplicity to be the same for all branches) and with n the order number of the node following the N^{th} symbol. Then Eq. (24) can be rewritten in the form

$$\log P_n = \log K_2 + \sum_{k=1}^{n} \lambda_k , \qquad (25)$$

where

$$\lambda_k = \sum_{j=(k-1)N_b+1}^{k\,N_b} \left[\log \frac{P(y|x)}{P_o(y)} - R\right]_j \tag{26}$$

is the contribution to the summation in Eq. (24) of the k^{th} branch examined by the decoder. Finally, we can drop the constant from Eq. (25) and focus our attention on the sum

$$L_n = \sum_{k=1}^{n} \lambda_k \quad , \tag{27}$$

which increases monatonically with the probability P_n.

A typical behavior of L_n as a function of n is illustrated in Fig. 6. The value of λ_k is normally positive in which case the probability that an error has been committed at some particular node is greater than the probability than an error has been committed at the preceding node. Thus, if the decoder has reached the n^{th} node and the value of λ_{n+1}, corresponding to the a posteriori most probable branch stemming from it, is positive, the decoder should proceed to examine the branches stemming from the following node on the assumption that the path is correct up to that point. On the other hand, if the value of λ_{n+1} is negative, the decoder should assume that an error has occurred and examine other branches stemming from preceding nodes in order of relative probability.

SECTION IX

A SPECIFIC DECODING PROCEDURE

It turns out that the process of searching other branches can be considerably simplified if we do not insist on searching them in exact order of probability. A procedure is described below in which the decoder moves forward or backward from node to node depending on whether the value of L at the node in question is larger or smaller than a threshold T. The value of T is increased or decreased in steps of some appropriate magnitude T_o as follows. Let us suppose that the decoder is at some node of order n, and that it attempts to move forward by selecting the most probable branch among those not yet tried. If the resulting value of L_{n+1} exceeds the threshold T, the branch is accepted and T is reset to the largest possible value not exceeding L_{n+1}. If, instead, L_{n+1} is smaller than T, the decoder rejects the branch and moves back to the node of order n-1. If $L_{n-1} \geq T$, the decoder attempts again to move forward by selecting the most probable branch among those not yet tried, or, if all the branches stemming from that node have already been tried, it moves back to the node of order n-2. The decoder moves forward and backwards in this manner until it is forced back to a node for which the value L is smaller than the current threshold T.

The implication of the decoder being forced back to a node for which L is smaller than the current threshold is that all the paths stemming from that node contain at least a node for which L falls below the threshold. This situation may arise because of a mistake at that node or at some preceding node, as illustrated in Fig. 6 by the first curve branching off above the correct curve. It may also result from the fact that, because of unusually severe channel disturbances, the values of L along the correct path reach a maximum and then

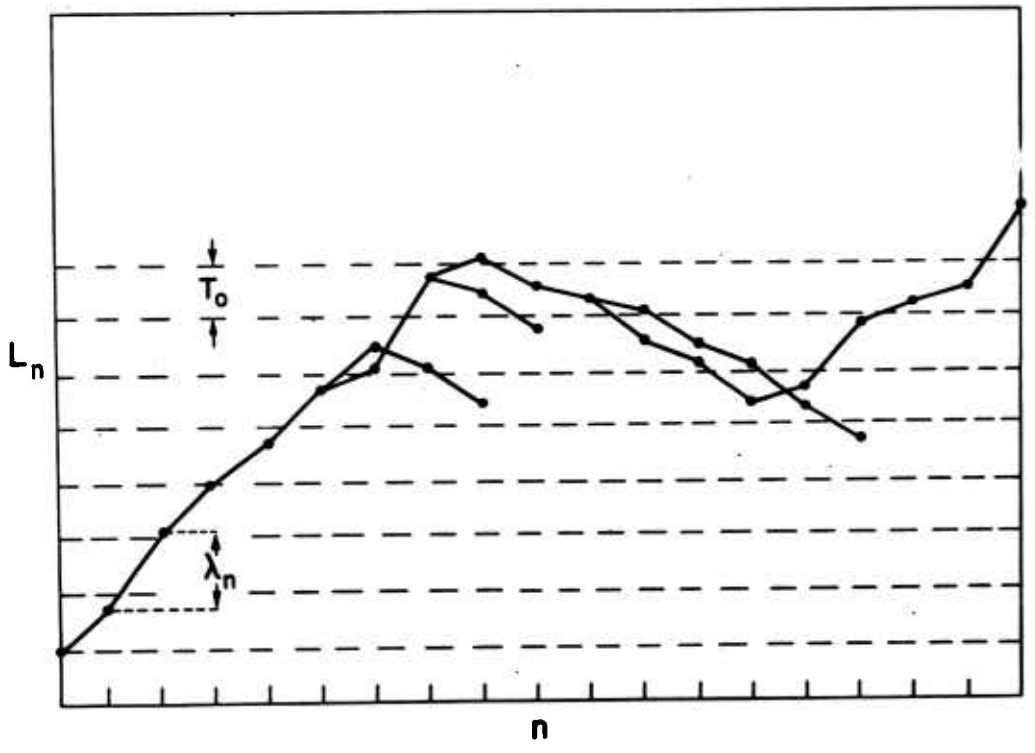


Fig. 6 Behavior of the Likelihood Function L_n Along Various Tree Paths. The Continuous Curve Corresponds to the Correct Path.

decrease to a minimum before rising again, as illustrated by the main curve in Fig. 6. In either case the threshold must be reduced by T_o in order to allow the decoder to proceed.

After the threshold has been reduced, the decoder attempts again to move forward by selecting the most probable branch just as if it had never gone beyond the node at which the threshold had to be reduced. This leads the decoder to retrace all the paths previously examined to see whether L remains above the new threshold along any one of them. Of course, T can not be allowed to increase while the decoder is retracing any one of these paths, until it reaches a previously unexplored branch. Otherwise, the decoder would keep retracing the same path over and over again.

If L remains above the new threshold along the correct path, the decoder will be able to continue beyond the point at which it was previously forced back, and the threshold will be permitted to rise again as discussed above. If instead L still falls below the reduced threshold at some node of the correct path or an error has occurred at some preceding node for which L is smaller than the reduced threshold, the threshold will have to be further reduced by T_o. This process is continued until the threshold becomes smaller than the smallest value of L along the correct path, or smaller than the value of L at the node at which the mistake has taken place.

The flow chart of Fig. 7 describes the procedure more precisely than it can be doen in words. Let us suppose that the decoder is at some node of order n. The box at the extreme left of the chart examines the branches stemming from that node and selects the one which ranks i^{th} in order of decreasing a posteriori probability*. Next, the value L_{n+1} is computed by adding L_n and $\lambda_{i(n)}$. The

*The value of λ for this branch is indicated in the chart by the subscript i(n), and the integer i(n) is assumed to be stored for future use for each value of n. The number of branches is $b = 2^{\nu_o}$. Thus $1 \leq i(n) \leq b$.

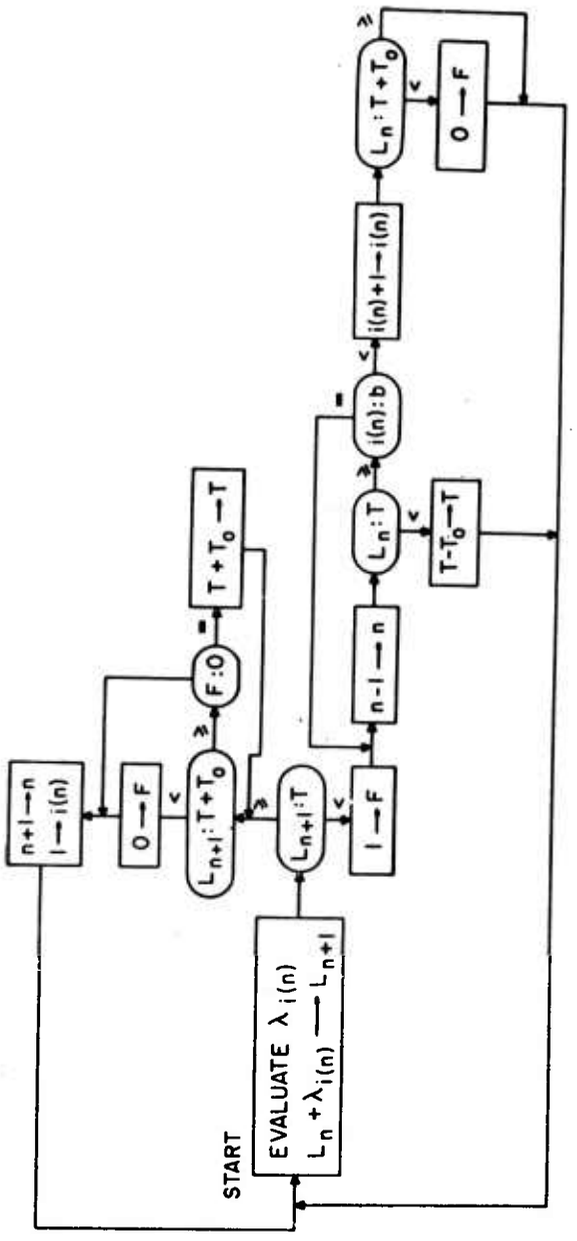


Fig. 7 Flow Chart of the Sequential-Decoding Procedure

value of L_n may be needed later if the decoder is forced back to the n^{th} node, and therefore it must be stored or recomputed when needed. For the sake of simplicity, the chart assumes that L_n is stored for each value of n.

The chart is self-explanatory beyond this point except for the function of the binary variable F. This variable is used to control a gate which allows or prevents the threshold from increasing depending on whether $F = 0$ of $F = 1$ respectively. Thus, F must be set equal to 0 when the decoder selects a branch for the first time, and equal to 1 when the branch is being retraced after a reduction of threshold. The value of F is set equal to 1 each time a branch is rejected; it is reset equal to 0 before a new branch is selected only if $T \leq L_n < T + T_o$ for the node to which the decoder is forced back. The value F is reset equal to 0 after a branch is accepted if $T \leq L_{n+1} < T + T_o$ for the node at which the branch terminates. It can be checked that, after a reduction of threshold, F remains equal to 1 while a path is being retraced, and it is reset equal to 0 at the node at which the value of L falls below the previous threshold.

SECTION X

EVALUATION OF THE PROCEDURE

The performance of the sequential decoding procedure outlined in the preceding section has been evaluated analytically for all discrete channels without memory. The details of the evaluation and the results will be presented in a forthcoming paper. The general character of these results and their implications are discussed below. The most important characteristics of a decoding procedure are its complexity, the resulting probability of error per digit, and the probability of decoding failure. We shall define and describe these characteristics in order.

The notion of complexity actually consists of two related but separate notions: the amount of equipment required to carry out the decoding operation, and the speed at which the equipment must operate. Inspection of the flow chart shown in Fig. 7 indicates that the necessary equipment consists primarily of that required to generate the possible channel inputs, namely a replica of the encoder, and that required to store the channel output and the information digits decoded. All other quantities required in the decoding operation can be either computed from the channel output and the information digits decoded, or stored in addition to them if this turns out to be more practical. We saw in Section I that the complexity of the encoding equipment increases linearly with the encoder memory ν, since the binary encoder must convolve two binary sequences of lengths proportional to ν. The storage requirements will be discussed in conjunction with the decoding failures.

The speed at which the decoding equipment must operate is not the same for all its parts. However, it seems reasonable to measure the required

speed in terms of the average number $\bar{n}$ of branches that the decoder must examine per branch transmitted. A very conservative upper bound to $\bar{n}$ has been obtained which has the following properties. For any given discrete channel without memory there exists a maximum information transmission rate for which the bound to $\bar{n}$ remains finite for messages of unlimited length. This maximum rate is given by

$$R_{comp} = \max_{P_o(x)} \left\{ -\log \sum_Y \left[\sum_X P_o(x) \sqrt{P(y \mid x)} \right]^2 \right\} . \tag{28}$$

Then, for any transmission rate $R < R_{comp}$, the bound on $\bar{n}$ is not only finite but also independent of ν. This implies that the average speed at which the decoding equipment has to operate is independent of ν.

The maximum rate given by Eq. (28) bears an interesting relation to the exponential factor α in the bound, given by Eq. (9), to the error probability for optimum block decoding. As shown in Fig. 4, the curve of α versus R coincides, for small values of R, with a straight line of slope -1. This straight line intersects the R axis at the point $R=R_{comp}$. Clearly $R_{comp} < C$. The author doesn't know of any channel for which R_{comp} is smaller than $1/2\ C$, but no definite lower bound to R_{comp} has yet been found.

Next, let us turn our attention to the two ways in which the decoder may fail to reproduce the information digits transmitted. In the decoding procedure outlined above no limit is set on how far back the decoder may go in order to correct an error. In practice, however, a limit is set by the available storage capacity. Thus, decoding failures will occur whenever the decoder proceeds so far along an incorrect path that, by the time it gets back

to the node where the error was committed, the necessary information has already been dropped from storage. Any such failure is immediately recognized by the decoder because it is unable to perform the next operation specified by the procedure.

The manner in which such failures are handled in practice depends on whether or not a return channel is available. If a return channel is available, the decoder can automatically ask for a repeat.[10] If no return channel is available, the strean of information digits must be broken into segments of appropriate length and a fixed sequence of $\nu - \nu_o$ digits must be inserted between segments. In this manner, if a decoding failure occurs during one segment, the rest of the segment will be lost but the decoder will start operating again at the beginning of the next segment.

The other type of decoding failure consists of digits erroneously decoded which can not be corrected regardless of the amount of storage available to the decoder. These errors are inherently undetectable by the decoder, and therefore, do not stop the decoding operation. They arise as follows.

The decoder, in order to generate the branches that must be examined feeds the information digits decoded to a replica of the encoder. As discussed in Section VI, the set of branches stemming from a particular node is specified by the last $\nu - \nu_o$ information digits. Then, let us suppose that the decoder is moving forward along an incorrect path and that it generates, after a few incorrect digits, a sequence of $\nu - \nu_o$ information digits which happen to coincide with those transmitted. This is a very improbable event because the decoder is usually forced back long before it can generate that many digits. However, it can indeed happen if the channel disturbance is sufficiently severe during the time interval involved. After such an event, the replica of the encoder (which generates the branches to be examined) becomes completely

free of incorrect digits, and therefore the decoding operation proceeds just as if the correct path had been followed all along. Thus, the intervening errors will not be corrected. As a matter of fact, if the decoder were forced back to the node where the first error was committed, it would eventually take again the same incorrect path.

The resulting probability of error per digit decoded is bounded by an expression similar to Eq. (9). However, the exponential factor α is larger than for block encoding, although of course it vanishes for $R = C$. This fact may be explained heuristically by noting that the dependence of the encoder output on his own past extends beyond the symbols corresponding to the last ν information digits. Thus, we might say that, for the same value of ν, the effective constraint length is larger for sequential encoding than for block encoding.

Finally, let us consider further the coding failures mentioned above. Since these decoding failures result from insufficient storage capacity, we must specify more precisely the character of the storage device employed. Suppose the storage device is capable of storing the channel output corresponding to the last n branches transmitted. Then, a decoding failure occurs whenever the decoder is forced back n nodes behind the branch currently transmitted. This is equivalent to saying that the decoder is forced to make a final decision on each information digit within a fixed time from their transmission. Any error in this final decision, other than errors of the type discussed above, will stop the entire decoding operation. No useful bound could be obtained to the probability of occurrence of the decoding failures resulting from this particular storage arrangement.

Next, let us suppose that the channel output is stored on a magnetic tape, or similar buffer device, from which the segments corresponding to

successive branches can be individually transferred to the decoder upon request. Suppose further that the internal memory of the decoder is limited to n branches. Then, a decoding failure occurs whenever the decoder is forced back n branches from the furthest one ever examined, regardless of how far back this branch is from the one currently transmitted.

Let us indicate with k the order number of the last branch dropped from the decoder's internal memory. There are two distinct situations in which the decoder may be forced back to this branch after having examined a branch of order k + n. The value of L along the correct path falls below L_k at some node of order equal to, or larger than, k + n; or it falls below some threshold $T \leq L_k$ at some earlier node, and there exists an incorrect path, stemming from the node of order k, over which the value L remains above T up to the node of order k + n .

An upper bound to the probability of occurrence of these events can be readily found. It is similar to Eq. (9), with $\nu = n\nu_o$, and a value of α approximately equal to that obtained for threshold block decoding.

SECTION XI

CONCLUSIONS

The main characteristic of sequential decoding that makes it particularly attractive in practice is that the complexity of the necessary equipment grows only linearly with ν, while the required speed of operation is independent of ν. Thus, it is economically feasible to use values of ν sufficiently large to yield a negligibly small probability of error for transmission rates relatively close to channel capacity.[6]

Another important feature of sequential decoding is that its mode of operation depends very little on the channel characteristics, and therefore, most of the equipment can be used in conjunction with a large variety of channels.

Finally, it should be stressed that sequential decoding is in essence a search procedure of the hill-climbing type. It can be used in principle to search any set of alternatives represented by a tree, in which the branches stemming from different nodes of the same order are substantially different from one another.

REFERENCES

1. W. W. Peterson, Error-Correcting Codes, M. I. T. Press and Wiley, 1961.

2. R. M. Fano, Transmission of Information, M. I. T. Press and Wiley, 1961.

3. C. E. Shannon, "A Mathematical Theory of Communication," Bell System Tech. J., 27, 379, 623 (1948).

4. J. M. Wozencraft, "Sequential Decoding for Reliable Communications," Technical Report No. 325, Research Laboratory of Electronics, M. I. T., 1957 — See also: J. M. Wozencraft and B. Reiffen, Sequential Decoding, M. I. T. Press and Wiley, 1961.

5. B. Reiffen, "Sequential Encoding and Decoding for the Discrete Memoryless Channel," Technical Report No. 374, Research Laboratory of Electronics, M. I. T., 1960.

6. K. M. Perry and J. M. Wozencraft, "SECO: A Self Regulating Error Correcting Coder-Decoder," IRE Trans. IT-8, No. 5, S128, Sept. 1962.

7. J. Ziv, "Coding and Decoding for Time-Discrete Amplitude Continuous Memoryless Channels," IRE Trans. IT-8, No. 5, 199, Sept. 1962.

8. R. G. Gallager, "Low Density Parity-Check Codes," IRE Trans. IT-8, 21, Jan. 1962.

9. R. G. Gallager, "Sequential Decoding for Binary Channels with Noise and Synchronization Errors," Report 25G-2, Lincoln Laboratory, M. I. T., 1961.

10. J. M. Wozencraft and M. Horstein, "Coding for Two-Way Channels," Information Theory, Fourth London Symposium (Edited by C. Cherry) Butterworth, London, 1961, p. 11.

RECENT CONTROL SYSTEMS THEORY

John G. Truxal*

SECTION I

INTRODUCTION

The recent, explosive growth of control technology (and the related control theory) can be traced to three somewhat diverse causes:

(1) The increased emphasis on military command and control systems in conjunction with the accelerating utilization of computer control in industrial automation, motivated by economic and performance considerations.

(2) The breadth of control engineering, with the increasing scope of meaningful applications in such directions as economic system analysis, management science, medical engineering, and societal engineering.

(3) The development of computer technology, with the consequent radical changes in the scope of engineering analysis and design and in the constraints on the class of useful and useable engineering systems.

This growth of the control and systems engineering field is mirrored by the intensive Russian and American research effort in industrial and nonprofit laboratories and in universities. **

* Polytechnic Institute of Brooklyn.

** In U. S. universities alone, the research budget approaches ten million dollars per year, according to the 1961 survey of the American Automatic Control Council.

In the following paragraphs, an attempt is made to indicate the status of current control theory and research: the nature of the problems under consideration, the extent to which theory relates to engineering practice, and certain directions particularly promising for future developments. Emphasis is directed toward those aspects of the theory which either have yielded interesting engineering results or promise such a yield in the near future; thus, in these pages the construction of a future systems theory, in the context of Dr. Drenick's remarks, is only an objective very secondary to the development of the portions of that theory which are needed for the solution of pressing current problems.

SECTION II

A MODERN CONTROL SYSTEM

Many of the requirements imposed on control theory can be illustrated by the relatively simple, specific example shown in Fig. 1: the control of a short-range vehicle by an electromechanical system which attempts to guide the vehicle along a line-of-sight path established by a human operator tracking the desired destination. The complete system thus involves two primary subsystems: the tracking loop and the guidance-control loop. (The system shown represents only one half of the overall configuration, which includes control of both the horizontal and vertical positions.)

In other words, as the target moves, the operator tracks with the optical-mechanical system in an attempt to hold the crosshairs on the target at all times. Simultaneously, the position of the vehicle is automatically compared with the crosshair position; the resulting error is modified in a computer to generate a control signal in coded form suitable for transmission to the vehicle. Within the vehicle, this control signal is used to actuate the corrective, control system which attempts to modify vehicle position in such a manner as to reduce the error toward zero.

The theoretical design problem for such a system consists of the following steps:

(1) Consideration of system structure.
(2) Simplification of the model.
(3) Analysis of performance.
(4) Design (in this case, of the computer program).
(5) Evaluation of performance.
(6) Introduction of complexities neglected in initial design.

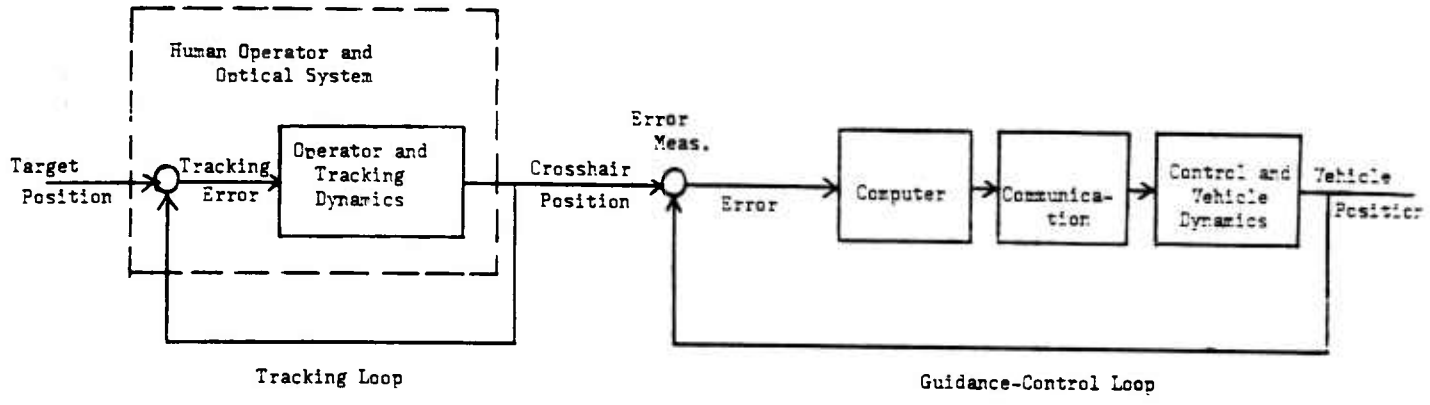


Fig. 1 Typical Control Problem

Obviously, engineering design seldom involves a steady progression through these steps, but rather normally entails a variety of feedback loops. Even in the simple form presented here, however, the design problem illustrates many of the elements of the modern engineering theory. *

First, the essential structure of the system is relatively complex (at least in comparison with the systems customarily described in the technical literature). The two-dimensional feature of the problem, involving the horizontal and vertical modes intercoupled by the nature of the vehicle dynamics and often also by the nature of the control-signal constraints, results in an overall system which can only be analysed with the use of analog or digital

---

* The selection of a simple guidance-control system as a vehicle for development of the thoughts of the following pages demands a certain justification, in the light of the central theme throughout these contributions of the information system sciences. Certainly to a considerable degree, an example such as Fig. 1 lies directly within the province of the control engineer, rather than the systems engineer.

The distinctions between this feedback system and the more grandiose systems for the detection and evaluation of enemy attacks (for example) are perhaps not as profound as might appear from superficial consideration. Certainly the contribution of the control and feedback engineer toward the development of an information system science must be based on the extension of the basic control theory to encompass systems which possess digital data processors rather than electromechanical equipment as components. This development from the specific and simple toward the general and complex has commonly characterized basic engineering research in the U.S.; it seems reasonable that information system science will evolve from the combination of such a direction and the merging of fundamental concepts from newly related scientific fields.

Thus, here emphasis is directed toward those aspects of feedback and control theories which relate to a broader system theory.

equipment. The determination of an approach to the design of such a system is still a problem which is largely unsolved (in spite of the extensive research efforts on multidimensional system analysis[1]) since, in situations such as the present, complete separation of the two outputs by noninteracting controls seems clearly incompatible with the demands for system economy, simplicity, and reliability.

Even if attention is focussed on the single-dimensional system of Fig. 1, difficult questions arise in connection with the desirability of the selected configuration. For example, the control system engineer usually bypasses the fundamental question: does the configuration represent an efficient utilization of the human operator? In our configuration, the operator is asked only to track the target with the crosshairs; the guidance-control loop does not utilize the prediction abilities, the learning abilities, or the adaptability of the human operator (e.g., his ability to compensate intelligently for certain equipment malfunctions or his adeptness in precognitive tracking). The desirable use of the human being depends in general upon such factors as the nature of the other tasks concurrently assigned to him, the environmental conditions, the degree of training which can be attributed to him, and the probability distribution of target motions.[2]

Consideration of the considerable and unusual abilities of the human operator to adapt and learn indicates, however, that in general the two loops of Fig. 1 should not be designed independently, that the human being should be allowed to influence directly the computer output. The additional performance characteristics achievable by a feed-forward transmission path from the human operator directly into the computer can be utilized to simplify computer design, over-ride certain malfunctioning of the other equipment, and improve overall system performance in even relatively simple cases. Even if this modification

of the configuration of Fig. 1 is not inserted, the two loops will be coupled to at least a minor extent if the operator can see not only the target but also the vehicle, since the operator will be aware of the overall purpose of the system and will inevitably attempt to abet the action of the guidance-control loop.

Inspection of Fig. 1 suggests immediately that improved performance could be achieved if a signal transmission path were added directly from the input target position into the computer. In the usual implementation, however, there is no possibility of generation of a signal measuring target position. The human operator can be used to develop a signal which is a function of the tracking error; alternatively, the crosshair-position signal can be modified dynamically (e.g., so that the input to the guidance-control loop contains a component predicting the future crosshair position to offset lags in both loops).

Thus, the choice of configuration,[3] in even the elementary problem considered here, is by no means simple and straightforward; an intelligent design decision can only be made on the basis of detailed studies of the man-machine system, and unfortunately, really only after analysis and initial design of the separate loops represented in Fig. 1 (i.e., after the completion of all six steps listed above). In actual situations the system engineer, harassed by pressing time schedules and confused by inadequate or insufficient data describing system components, customarily selects a configuration rather arbitrarily (but with finality) and proceeds as rapidly as possible to the better-defined problem of the design of a specific system component (in the case of Fig. 1, the computer program).

This brief description above of the difficulties arising at the outset of system design is included here to emphasize the large degree of arbitrariness which characterizes so many control system problems — an arbitrariness which tends to be forgotten once we are immersed in the details of computer design

and the problems (for example) of optimizing the guidance-control loop. Before considering the nature of this more detailed loop design, however, we mention briefly certain additional aspects of the system design represented in Fig. 1 — aspects which further characterize so many of the applications of modern control theory.

In addition to the configuration complexity and the involvement of a human operator, design is often complicated by the time-varying nature of the control and vehicle dynamics. For example, if the vehicle changes speed markedly during the duration of the control interval, the "natural frequencies" may vary by large factors over time intervals only slightly larger than the system response time. In such a situation, the concepts of natural frequencies, damping ratios, and transfer functions can not be interpreted, and analysis requires a return to the simultaneous differential equations derived from the basic physics underlying the operation of the system components.

An additional complication is introduced by the existence of time delays in one or more parts of the system. In our specific example of Fig. 1, significant time delays may be present in the action of the human operator (unless the target motion is sufficiently simple to permit the operator to anticipate future values), in the computer (because of the time required to code, to solve the equations of control, and to generate the coded computer output signal), and in the transmission system.

The existence of ideal time delays in an analog system results in a mathematical model involving differential difference equations; the resulting analysis of even relatively simple systems is markedly more difficult than with the conventional analog or digital system. While there seems to have been a steady increase in the Russian literature on the problem of the analysis of feedback systems with pure delays (particularly with emphasis on process

control problems in which delay results from the time required for transportation or for chemical or physical reactions), there seems to be a dearth of meaningful approaches to the resulting problems of system analysis. Perhaps the most useful approach involves a return to the classical time-domain model, with a sampler inserted artificially to convert the mathematical model to a set of difference equations. (If the sampling period is chosen so that the delay is an integral multiple of this period, the delay is simply represented mathematically.)

For example, the determination of the unit impulse response for the simple configuration of Fig. 2, including a process with a delay of one second, involves the time-domain solution of the relation

$$\left[u_o + k(t) \right] \; e(t) = r(t) \quad , \tag{1}$$

where u_o is the unit impulse, k the impulse response for $K(s)$, e the error, and r the input (a unit impulse). If we choose a sampling period of one second, and if $k(t)$ is then given by the sequence of sample values

$$\left\{ k \right\} = 0,\ 0,\ 3/2,\ 7/4,\ 15/8,\ 31/16,\ \ldots \tag{2}$$

a straightforward numerical solution of Eq. (1) yields an e sequence directly by long division:

$$\left\{ e \right\} = 1,\ 0,\ -3/2,\ 7/4,\ 3/8,\ -31/16,\ \ldots \tag{3}$$

In the absence of the one-second delay, this unstable response is stable and exceedingly simple:

$$\left\{ e \right\} = 1,\ -3/2,\ 1/2,\ 0,\ 0,\ 0,\ \ldots \tag{4}$$

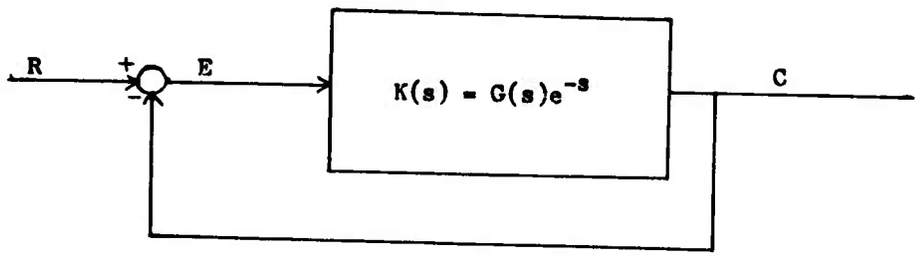


Fig. 2 Feedback System with Unity Delay

A straightforward and conceptually simple analysis of this type permits evaluation of the response of actual systems of considerable complexity. The existence of time delay is largely immaterial in its effect on analysis complexity. Unfortunately, such numerical analysis (or analysis with discrete state system theory) is only somewhat indirectly useful in system design, primarily because the analysis does not lead in general to global solutions.* The situation here is directly comparable to that which exists when we compare numerical analysis methods with the differential and integral calculus: while discrete system analysis clarifies certain fundamental concepts of the calculus and permits the solution of specific problems of awesome complexity, the calculus leads naturally to global solutions and general interpretations.

Thus, the modern control systems design problem very often involves a complex system configuration (including a number of interlocking feedback loops and numerous inputs and outputs), elements which are time-varying or nonlinear or which include a human operator, various time delays which may vary during operation, and finally (although not mentioned above) elements which are only very vaguely understood and poorly characterized mathematically.

---

* We certain can derive stability tests, for example, for systems described by sets of linear difference equations with constant coefficients. In the analysis of complex systems, however, we in general do not have available simple techniques for the determination of the influence of a particular parameter on system performance. The problem of interpreting the numerical analysis is further complicated by uncertainty as to how to select the sampling period (in the analysis above, for example). We would like to select the sampling period as large as possible to simplify calculations, but sufficiently small to insure our discrete model yields a response adequately close to the actual response of the continuous model. Again, no simple theory exists for determination of the compromise.

System design involves, correspondingly, system studies to arrive at some sort of decision on the selection of a configuration — a decision which usually must be reached relatively early in the design procedure (whether we are discussing a specific and relatively simple control problem as protrayed in Fig. 1, or we are concerned with an information processing system involving primarily communication links and computers for the data storage and the decision functions — the problems seem to differ primarily in the nature of the components, rather than in the philosophy of design).

The aspects of system design discussed above are essentially concerned with what we might term Stage 1 of the system design problem: the transition from the customary, vague statement of broad system objectives to a configuration and a tractable model for the various elements of the system. Stage 2 is concerned with mathematical design of the "free" elements (e.g., the computer of Fig. 1) — the components in which at least certain of the parameters can be adjusted within specified bounds. In the following section, we consider the status of modern control theory for the second stage, which essentially completes the theoretical aspect of the system design. [In an actual problem, these two stages are followed, of course, by specific realization of desired component characteristics (frequently a design problem in itself), prototype construction and test and/or computer simulation system studies (for reliability and life-test evaluations, for example), actual system testing and evaluation, etc.]

SECTION III

OPTIMIZATION THEORY

The central theme of control and feedback systems research at the present time is unquestionably optimization: the design of physical systems to yield performance which is analytically optimum according to a selected performance criterion. While the control systems engineer hopes to apply this optimization theory to the design of the overall system (as depicted in Fig. 1, for example), mathematical and conceptual difficulties ordinarily require that the optimization theory be applied to the design of specific elements such as the computer, in order to yield an optimum within the constraints imposed by the arbitrary selection of configuration and the simplification of the models of the process being controlled, the human operator, etc.

Optimization theory, as we shall briefly describe it in the following paragraphs, is motivated by:

(1) The difficulties which arise when we attempt to apply conventional control theory to complex problems such as the system of Fig. 1. In even moderately complicated situations, the designer is confronted with the difficult question of how to start. Optimization theory delineates the necessity for selection of a mathematical performance criterion, the type of models required for characterization of the components, and the form of an appropriate control system (i.e., the solution to the optimization problem).

(2) The need for an "absolute" basis for the evaluation of systems designed by other approaches, including empirical methods.

(3) The possibility of realizing useful solutions. For example, in a variety of idealized guidance problems, optimization theory has led to such significant improvements in system performance that there can be no question of the significance of the solution even if second-order effects were included.[4] A problem falling into this category is the determination of a program of attitude and velocity versus altitude to bring a piloted aircraft from the ground to a pre-specified altitude and horizontal flight speed in minimum time.

The excitement of optimization theory resides in the strong promise that significant results will be realized in all three of these directions.

Any discussion of modern optimization theory is difficult because the more general we attempt to be in our theory, the more simplifications are necessary to bring the problem within tractable bounds. While this conflict exists throughout engineering, it is particularly troublesome in optimization theory, since there is a fundamental question of the value of an optimum solution for an idealized and hypothetical problem: is this "optimum" solution of the hypothetical problem actually better than a non-optimum solution of the real problem?

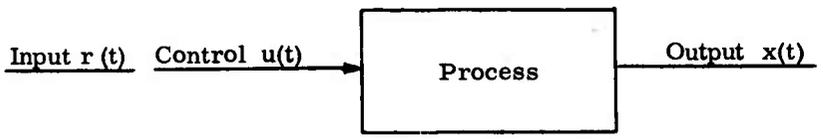


Fig. 3 Optimization Problem

The difficulty can be illustrated with reference to Fig. 3, which indicates the notation used below. The control problem is: given the reference input r and the process and the desired behavior of the response x, how do we determine

the control signal u ? The optimization problem is straightforward if we can impose adequate constraints on the various elements of the problem:

(1) The signals — for example, if r is known to be a member of a narrow class, such as the polynomials of second order in t, or if r can be approximated by such a function over intervals of time significantly larger than the system response time.

(2) The process — if the process is completely known.

(3) The state of the system — if the present state is known or immediately measurable.

(4) The criterion — if the performance criterion to be maximized or minimized is specified.

(5) The constraints — if the constraints are, for example, saturation limits on the components of u, and do not involve constraints on the state components.

(6) The disturbances — if there are no disturbances influencing the future state of the system, or if the effects of such disturbances can be determined precisely.

(The above are not necessary conditions for the solution of the optimization problem, but rather are listed to indicate the complications which arise when we attempt to apply the theory in real-life situations.)

If all of the conditions cited above are applied, the resulting control problem is hardly exciting. Indeed, in the challenging systems problems, we find that several, if not all, of the conditions are violated. Furthermore, in actual situations the goal of system design if frequently improved performance,

rather than optimum performance in the sense of the minimization of some arbitrarily selected, mathematically convenient integral function of the system error.

As a consequence of these difficulties, and in the light of the demands on the theory as expressed at the beginning of this section, two quite different approaches to optimization have developed in the last few years in the control systems field: approaches which we term here restricted optimization and general optimization, although these names perhaps imply an unjust relative evaluation.

SECTION IV

RESTRICTED OPTIMIZATION

Within the realm of restricted optimization, we include those design approaches which focus on the concept of improved system performance, and which attempt to realize this goal by restricting consideration to a sub-optimum problem of system design. For example, within this category fall those systems which operate on the basis of the adjustment of a single parameter (or a small number of parameters) to extremalize a selected function of system performance. The system configuration and the individual elements are selected according to conventional control techniques, after which the optimization is included in order to improve performance within the framework of the original system design.

The technical literature of the feedback control is replete with examples of such sub-optimal systems developed during the last few years. In general, such designs are characterized by simplicity both conceptually and practically, reliability, and emphasis on practicality. In terms of control systems theory, such designs represent major contributions primarily in terms of the novel configurations which result — systems which never would evolve from the conventional control theory focussing so heavily on stability considerations for linear, single-loop feedback configurations.

The research efforts along this direction of restricted optimization can be illustrated by three specific approaches, the first of which is depicted in Fig. 4, a sketch of the M. I. T. Instrumentation Laboratory adaptive system which has evolved from the early work by Draper and Li on optimalization. In this configuration, a conventional feedback control system (represented in simplified form by the single control loop in the figure, even though in most

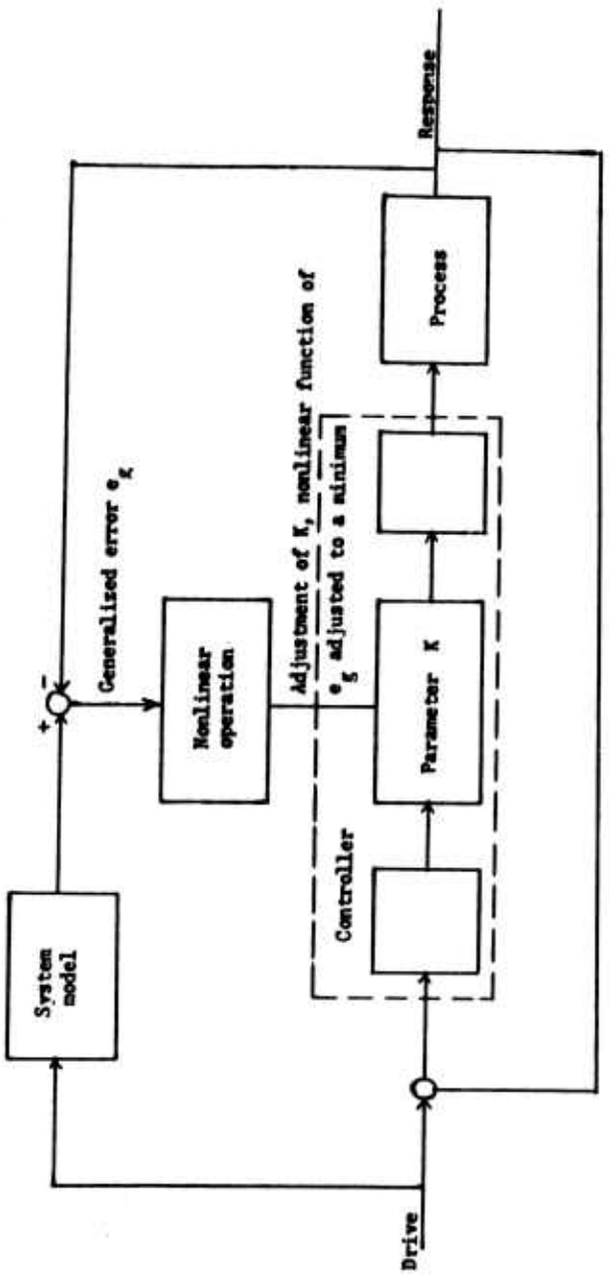


Fig. 4 Model-Reference Adaptive System

cases the configuration is multi-loop) is designed; superimposed on this is the optimizing subsystem. The actual system output is compared with the response of a model to yield a generalized error e_g. A nonlinear function of this e_g is used to adjust a parameter K of the control loop to yield a minimum of the performance criterion.

Certain basic problems arise in the design of such a "model-reference adaptive system:"

(1) How is the model to be chosen?

(2) How are we to select that parameter K which is to be varied?

(3) How is the criterion function chosen?

(4) Under what conditions is the optimizing loop itself stable?

While certain aspects of these problems have been investigated, * the extension of this approach to complex configurations in the sense of Section II of this paper relies heavily on the final design and verification of performance characteristics via simulator studies. Such a situation should be neither surprising nor discouraging, however, in view of our earlier thought that optimization techniques are primarily useful in the design of complex systems in which the engineer is faced with the difficult question of how to start.

The extensive development of the configuration of Fig. 4 has been motivated, at least in part, by the adaptive autopilot problem for piloted aircraft moving through radically varying environments — a specific engineering problem which also stimulated the development of the Minneapolis-Honeywell autopilot configuration.[7] Complementary to the military interest in optimization has been

---

* In particular, recent work has focussed on the selection of the model[5] and the relation of the sensitivity function to the selection of the parameter K.[6]

the emphasis on industrial applications of computer control; once the computer is installed for data logging and routine processing of system performance information, for shutdown, startup and emergency control, and for the automation of routine economic analyses, the systems engineer visualizes utilization of the computer flexibility to realize optimum plant performance, or at least control which is more precise and faster than that achievable with human operators.

In the technical literature describing applications of optimum computer control, two approaches dominate: the model approach and the automatic experimental approach.* Under the former design philosophy, a model of the process is used to determine the optimum control signal as a function of measurable signals and disturbances. The response can be very rapid, can avoid difficulties with multiple extrema, and can include learning or updating modification in the model; on the other hand, performance is limited by the accuracy of the model, and realization of performance near optimum requires intensive studies of and measurements on the process to be controlled.

In the automatic experimental approach, the optimum is realized by the injection of artificial input signals to perturb the operating point, with a subsequent evaluation of whether the performance improved or deteriorated as a result of the change. In such an approach, the speed of response is limited by the fact that the system must search for an optimum, the presence of multiple extrema causes difficulties, and the existence of many signal variables leads to exceedingly slow and possibly poor performance. A considerable portion of the optimization literature of the past few years is devoted to study of searching techniques to overcome one or more of these difficulties.

---

* The terminology here is that of the Chen-Decker article,[8] which also includes references to a number of actual applications of industrial computer control.

Chen and Decker[8] emphasize the advantages to be gained by combining these two approaches, with the performance of the composite system indicated in Fig. 5. The plot shows the payoff J as a function of the control signal u for various disturbance inputs (d_1 and d_2 specifically indicated). The solid curve portrays the variation of J with u for the actual plant, the dashed curve the corresponding solution for the simplified model of the plant. The two constraint curves indicate the allowable bounds on u and J which result from considerations such as safety or which are imposed in order to avoid subsidiary extrema.

If the system is initially operating at point 0 with a disturbance input d_1, and the dynamic performance is demonstrated by a change from d_1 to d_2, we find that the performance moves initially to point E since u can not change instantaneously. As fast as the model system responds, operation moves to point C;* thereafter the automatic experimental procedure moves the system toward the optimum operating point B. Clearly we are using the model here for fast, gross corrections, the automatic experimental system for the slower, fine corrections, although the specific inter-relation of these two portions of the system may be quite complicated in problems more realistic than the simple situation depicted in Fig. 5.

---

* The system was initially at 0, rather than the model value D. This offset from D was the result of the automatic experimental procedure. If this offset remains after the step change in d, the model action may bring operation to A rather than C; the actual location depends on the way the model and the experimental signals are inserted. In any case, however, motion toward B follows the action of the experimental equipment, and in many applications the details of motion are of little interest.

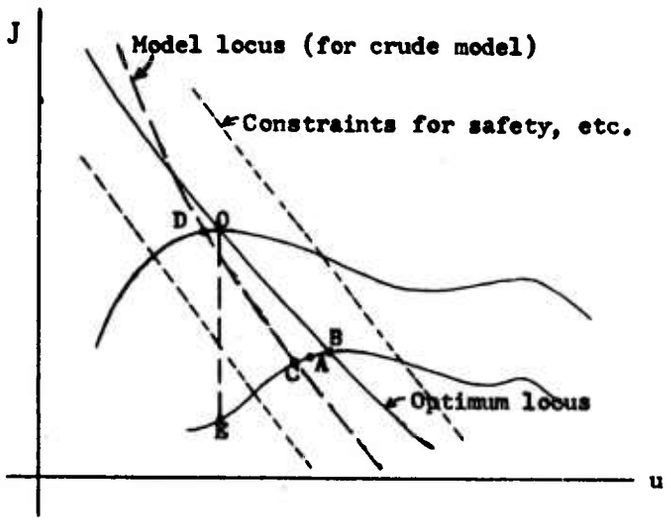


Fig. 5 Second Example of Restricted Optimization

The final aspect of restricted optimization to be discussed here is based upon the utilization of simple digital logic for the determination of the desired control signal u on the basis of inspection of the response of a high-speed model of the process. If for conceptual simplicity we assume the discrete case, we can consider in Fig. 4 the problem of selecting the control signal u which is to be constant over each interval of time. At the time $t=0$, we wish to determine the constant control signal for the first interval from 0 to T. This determination must be derived from the available information: the present and past values of the control signal u, the response x, and the reference input r.

The determination can be implemented in the following way. We consider only the values of r, x, and u every T seconds (i.e., at the sampling times). On the basis of the known statistical characteristics of r, we can estimate $r_1, r_2, r_3, \ldots$ (the values of r at $T, 2T, 3T, \ldots$). On the basis of the past values of u and x and our knowledge of the process dynamics, we can determine the response values into the future $(x_1, x_2, x_3, \ldots)$ for any assumed sequence of control signal values $u_0, u_i, u_2, \ldots$. If we wish to minimize the mean square error, we might attempt to select u_0 in such a way that (assuming u_1, $u_2, \ldots$ are later selected optimally) we would minimize the summation

$$(r_1 - x_1)^2 + (r_2 - x_2)^2 + \ldots$$

It seems apparent that the choice of u_0 has a decreasing effect on the successive terms in this series; furthermore, because of the increasing difficulty of predicting r_j as we look farther into the future, we should weight more heavily the early terms. Such considerations suggest the consideration of only a small number of these terms, for example three:

$$J = (r_1 - x_1)^2 + (r_2 - x_2)^2 + (r_3 - x_3)^2 \quad . \tag{5}$$

Here r_1, r_2, and r_3 are predicted values of r; the x_j are future system response values which incorporate the effects of past signals plus the effects of the variable or controllable future signals u_0, u_1, and u_2. The optimization involves determination of

$$J_{min} = \min_{u_0,\, u_1,\, u_2} \left[(r_1 - x_1)^2 + (r_2 - x_2)^2 + (r_3 - x_3)^2 \right] , \qquad (6)$$

but we actually shall apply to the process only the signal u_0. When $t = T$ (and u_1 is to be applied), we shall re-evaluate the new optimum value for the control signal.

The attractiveness of this approach to sub-optimization derives from the possibility of simplification in the evaluation of u_0. If we consider a binary control signal (u_0 equals -1 or +1), we must choose between these two values on the basis of minimization of J, where we assume u_1 and u_2 will subsequently be selected in an optimum manner. In terms of the logic required to implement this decision on u_0, we need to divide our three-dimensional space into two parts: one requiring $u_0 = +1$, the other $u_0 = -1$. Our present location in this three-dimensional space is determined from the three predicted values of the future system input and the response with all possible inputs.

Major further simplication is possible[9] if the process is linear, so that the future response can be divided into two, additive parts resulting from past inputs and from future inputs. Then our present location in this three-dimensional space is determined by the predicted values of the future system error with no control input. Implementation of this control scheme then involves only high-speed prediction of e_1, e_2, and e_3 without control, and then (by simple digital logic) determination of the location of this "state" with respect

to the division of "state" space into the two parts corresponding to the two possible control signals. Thus, this system involves only a high-speed model of the process to predict future response values from the present energy storage conditions, a predictor to act on the present and past input, and simple digital logic to determine in which half of the three-dimensional space we are situated.

The three schemes described in this section are only three of a wide variety of practical system realizations derived with emphasis on the sub-optimization problem. In each case, design of the optimizing equipment requires at least a reasonable estimate of process dynamics;* in each case, the optimizing system is designed to correct for slow process variations or the effects of low-frequency disturbance inputs; in each case, the optimizing components are sufficiently simple to permit simultaneous realization of control over reliability. These advantages are at least to some extent offset by the fact that meaningful analysis of the three systems is apparently not possible if we are working with anything other than the most elementary processes; actual design and verification of the value of optimization must rest upon computer simulation studies and actual equipment tests. This difficulty is perhaps more alarming to the professor than to the control system designer, however, since the latter seldom is concerned with systems amenable to detailed analysis, regardless of the design philosophy employed. Certainly in the three directions, control engineering is beginning to accumulate an impressive list of successful applications.

---

* For example, in Fig. 4 neither the model nor the parameter can be selected without this knowledge; both the last two methods depend upon a model of the approximate process characteristics, even though in both cases the model can be improved automatically during normal system operation.

SECTION V

GENERAL OPTIMIZATION

In the examples of the preceding section, the configuration is selected at the outset; the optimization considered is based upon variation of a single parameter (or a very few parameters) to realize the best possible performance from this configuration. In this section, we mention briefly the recent work on the more general problem of optimization of system performance, with optimization based only on the specified process and signals. As indicated in Section III, if we are to obtain solutions of this more general problem, we must accept additional constraints on the problem specification. For example, optimization procedures assume the plant dynamics are known entirely — precisely in the deterministic cases and in terms of the relevant statistics in the stochastic cases.

The types of optimization problems which have been investigated cover a wide range, depending on the hypotheses as to the nature of the known aspects of signals, process dynamics, and constraints. Two typical problems are:

(1) Minimization of the time required to reach a specified system state from the given initial state, with constraints on the control signal (e.g., the components of u subject to saturation). The work in this direction represents a generalization of the well-established analysis of the bang-bang problem.

(2) For control of a vehicle, minimization of the expected error at a certain point in the trajectory. In a common form, this problem clearly illustrates the three elements of the optimization problem: criterion, constraint, and conflict. The criterion of performance

is the expected error. The constraints are a specified, limited quantity of fuel, uncertainty in the effects of using a given amount of fuel, and uncertainty in the estimation of the present vehicle position and velocity. The conflict arises because the longer we wait before making a correction, the more accurate is our information on the vehicle state; on the other hand, we would like to make the corrections as early as possible in order to realize greater improvements in the final accuracy from a given quantity of fuel.

Thus, the specific optimization problem depends upon the criterion, the constraints, and the particular conflict involved, as well as on the specific plant dynamics and the known characteristics of the exciting signals.[10, 11, 12]

The form of the optimization problem can be illustrated in terms of an exceedingly simple example. If the process is described by the differential equation in vector form

$$\dot{x} = f(x, u) \ , \tag{7}$$

and by the initial state $x(t_o)$, and if the criterion is given by the integral

$$J = \int_{t_o}^{T} F(x, u)\, dt \ , \tag{8}$$

we seek the optimal control policy u_0 which minimizes (or maximizes) S. Clearly, two forms of solution are possible. In the simple case outlined above, the optimal control signal u_0 depends on $x(t_o)$ and t; in other words, at the outset we can find the control to apply throughout the time interval from t_o to T. In such a case, which describes the majority of recent optimization

work, the solution is open-loop: we do not need to measure continually $x(t)$, the state of the process, and the control signal does not depend upon process response.

While the justification for the use of feedback in optimal and nonlinear systems is certainly not theoretically apparent, we do feel intuitively that feedback should assist in the reduction of the effects of unmeasurable disturbances and uncertainties or variations in the process parameters. The close-loop solution can be realized if we can find a solution u_0 which depends upon continuing measurement of the system state. Alternatively, we can employ a sampled data approach, periodically measuring $x(t)$ to update our optimal control policy by a re-determination of the optimal design, with the measured $x(t)$ in the role of $x(t_o)$ initially, and the interval of integral in Eq. (8) changed correspondingly.

The two basic solutions to the optimization problem are provided by Pontryagin's maximum principle[10] and Bellman's dynamic programming.[13] In the former approach, we encounter the necessity for the solution of a nonlinear two-point boundary value problem, with the corresponding need for research in computer and simulation techniques for solution. The problem can be solved by arbitrarily assuming $x(T)$, rather than the given $x(t_o)$, then working backward in time, if we are able to scan until all values of $x(t_o)$ are obtained. If we want a feedback system operating on the basis of measurement of $x(t)$, we need computer storage able to handle the entire range of $x(t)$. In order to avoid the storage and computational difficulties, recent research emphasizes the application of steepest descent methods for evaluation of the optimal, open-loop control law.

In the dynamic programming approach, we calculate in reverse the optimum trajectories — again any meaningful problem leads to extreme computational and storage difficulties unless logical procedures can be determined for finding simple approximations to the optimal control policy. (In the dynamic programming approach, however, we are led logically to a feedback structure, since the solution indicates directly the manner in which updated and improved data on the system state should be used.)

SECTION VI

CONCLUSION

Modern feedback and control systems theory is being developed under the pressure of a wide range of complex systems problems arising from both military and industrial extensions of control technology. Faced by the myriad of system design problems in which a logical starting point is not even apparent, the control system scientist (and in particular, the applied mathematician) has focussed attention on optimization theory.

The recent work in optimization has not only led to promising guidelines for the solution of actual problems, but of perhaps even greater significance, this work has led to a number of fundamental concepts new in control technology. For example, the mathematical techniques of dynamic programming lead to new insight into system description, suggest ways to reformulate previously difficult problems, and indicate logical techniques for basic problems such as the measurement of system state. The intensive study of optimization by Kalman,[14] in particular, has led to definitions of controllability and observability, with novel results in such directions as filtering theory and our understanding of the relationship between various process models (e.g., the transfer function and the state description). Scientific results of this nature provide a foundation for a startling expansion of control system science and a gradual development from the specific problems of engineering research toward the general systems theory of the future.

REFERENCES

1. SIAM Symposium on Multivariable Linear Control System Theory, Cambridge, Mass., November 1-2, 1962.

2. George A. Bekey, "An Investigation of Sampled Data Models of the Human Operator in a Control System," Report ESD-TDR-62-36, Wright-Patterson Air Force Base, Ohio, May 1962.

3. W.A. Lynch and J.G. Truxal, "Concepts of Feedback Theory," MRI Technical Report, January 1961, Polytechnic Institute of Brooklyn, Brooklyn, N.Y.

4. Symposium on Optimizing Control Theory, Wright-Patterson Air Force Base, Ohio, September, 1962.

5. Y.T. Li and P. Whitaker, Some Research Work in Self Adaptive Systems in MIT Aeronautical and Astronautical Engineering Department, Paper presented at IFAC Symposium on Adaptive and Optimizing Control, Rome, Italy, April, 1962, to be published by the Instrument Society of America.

6. P. Whitaker, Model-Reference Adaptive Systems, Talk presented at the University of Florida Winter Workshop on Optimizing Control, February 22, 1962.

7. D.L. Mellen, Application of Adaptive Flight Control, Paper presented at IFAC Symposium on Optimizing and Adaptive Control, Rome, Italy, April, 1962, to be published by the Instrument Society of America.

8. K. Chen and R.O. Decker, Process Optimization by Combining the Model and Experimental Approaches, ISA Transactions, 3 (279-285, July, 1962.

9. S. Horing, On the Optimum Design of Predictor Control Systems, paper submitted for IFAC Second International Congress, Basle, Switzerland, September, 1963.

10. G. Leitmann, "Optimization Techniques," Academic Press, New York, N.Y., 1962.

11. R. E. Kalman, T. S. Englar, and R. S. Bucy, "Fundamental Study of Adaptive Control Systems," Technical Report ESD-TR-61-27, Vol. 1, Wright-Patterson Air Force Base, Ohio, April 1962.

12. B. Friedland, "Optimum Space Guidance and Control," Melpar Technical Note 62/3, Applied Science Division, Melpar Inc., Watertown, Mass., June 1962.

13. R. Bellman, "Adaptive Control Systems: A Guided Tour," Princeton University Press, Princeton, N.J., 1961.

14. R. E. Kalman, Canonical Structure of Linear Dynamical Systems, Proceedings of the National Academy of Sciences, 48-4, pp. 596-600, April 1962.